

Article

Confidence-Calibrated Consistency Matching for Semi-Supervised Image Classification Under Extreme Label Scarcity

Dong-Hyun Won ¹, Hyuk-Gyu Park ^{2,*},[†] and Kwang-Seong Shin ^{3,*},[†]¹ Office of Liberal Arts Education, Wonkwang University, Iksan 54538, Republic of Korea; dhwon79@wku.ac.kr² Department of Computer and Software Engineering, Wonkwang University, Iksan 54538, Republic of Korea³ Department of Computer Engineering, Suncheon National University, Suncheon 57992, Republic of Korea

* Correspondence: hgpark7@wku.ac.kr (H.-G.P.); waver@scnu.ac.kr (K.-S.S.)

[†] These authors contributed equally to this work.

Abstract

Labeling images is expensive, but unlabeled data is abundant. Semi-supervised learning (SSL) addresses this gap, though the dominant pseudo-labeling methods can suffer from confirmation bias—reinforcing their own confident-but-wrong predictions—most severely when labels are scarcest. Under a controlled, reproducible compute-constrained protocol on CIFAR-10, SVHN, and CIFAR-100, we examine which ingredients of consistency-based SSL actually help when as few as four labels per class are available. We propose CCM (Confidence-Calibrated Consistency Matching)—a per-class curriculum threshold, a dual strong-view consistency loss, and a smooth confidence weighting that softly admits borderline pseudo-labels—together with a unified view in which FixMatch, FlexMatch, SoftMatch, and CCM instantiate a single generalized weighting function. On CIFAR-10 with 40 labels, CCM reaches 35.14%, a significant improvement over the FlexMatch design it directly extends (+3.30 percentage points (pp), paired t -test $p = 0.001$). SoftMatch, re-trained under the identical budget, performs better still at the two smallest budgets (37.04% at 40 labels, $p = 0.049$), while CCM leads numerically at 4000 labels: the two smooth-weighting designs top the extreme-scarcity board—convergent evidence that the smoothness of the weighting function, more than the placement of its threshold, is the decisive design axis. We also report a negative result: cross-view agreement helps neither as an admission gate nor as reliability reweighting, reducing accuracy by up to 4.37 pp; agreement is a positive correctness signal, but its absolute level (approximately 44% correct among agreeing pseudo-labels) is too low to filter on safely. Curriculum thresholding, by contrast, hurts on the easier SVHN dataset and fails outright on CIFAR-100 when its per-class statistics become too thin. CCM adds no inference-time cost. We do not claim universality; we characterize when each ingredient helps within a single, identical-budget protocol.



Academic Editor: Manohar Das

Received: 22 June 2026

Revised: 27 July 2026

Accepted: 31 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: semi-supervised learning; label scarcity; pseudo-labeling; confirmation bias; consistency regularization; image classification; calibration; negative results

1. Introduction

Supervised deep learning owes much of its success to large annotated datasets, yet in many real-world settings, obtaining labels is slow, is costly, or requires scarce expert attention, while unlabeled images accumulate at negligible cost. In medical imaging, expert annotation remains the binding constraint on model development, and semi-supervised

training on unlabeled scans is an active remedy [1]; in remote sensing and industrial sensing applications, hyperspectral and inspection imagery is produced far faster than it can be annotated [2]. The resulting regime—a small labeled set together with a large unlabeled pool—is one of the canonical data-related challenges for machine learning and the target of semi-supervised learning (SSL); two peer-reviewed journal surveys [3,4] give a comprehensive overview of the field.

The most effective modern SSL methods for image classification combine consistency regularization with pseudo-labeling: a model's confident prediction on a weakly augmented image is used as a hard target for its prediction on a strongly augmented version of the same image. FixMatch [5] crystallized this recipe and remains a strong, simple baseline; FlexMatch [6], FreeMatch [7], and SoftMatch [8] refine how many and which pseudo-labels to trust through adaptive thresholds and soft weighting.

A persistent obstacle to all pseudo-labeling methods is confirmation bias [9]: once the model makes a confident but incorrect prediction on an unlabeled example, training on that pseudo-label reinforces the error, which raises the model's confidence, which admits the example past any fixed threshold, and so on. The effect is self-amplifying and is worst when labels are scarcest because the few labeled examples cannot adequately anchor the decision boundary.

Despite the breadth of this literature, two gaps remain. First, published SSL results are typically obtained under training schedules of a million or more optimization steps with per-method hyperparameter tuning, so the reported rankings entangle each method's mechanism with its tuning budget; statistically tested comparisons in which every method receives an identical, modest budget are rare. Second, the extreme end of label scarcity—around four labels per class—is exactly where confirmation bias is most damaging and where intuitively appealing remedies are least examined. This paper therefore asks a deliberately narrow, practical question:

Research question. Under extreme label scarcity and a fixed, modest compute budget, which ingredients of consistency-based SSL improve accuracy, which have no effect, and which are harmful?

We make three contributions, each mapped to this question.

1. We propose CCM (Confidence-Calibrated Consistency Matching), a combination of three previously studied ingredients: a per-class curriculum threshold, a dual strong-view consistency loss, and a smooth confidence weighting that admits borderline pseudo-labels gradually instead of applying a hard cutoff. We also give a unified formulation (Section 3.4) in which FixMatch, FlexMatch, SoftMatch, and CCM are instances of one generalized sample-weighting function; within this frame, the three ingredients address complementary failure modes, and CCM's specific choice—a continuous weighting centered on a class-conditional curriculum threshold—occupies a point in the design space that prior methods leave empty. In the extreme-scarcity regime that we focus on (40 labels on CIFAR-10), CCM improves significantly on FlexMatch, the hard-threshold curriculum method it directly extends ($p = 0.001$), and the two smooth-weighting designs—CCM and the re-trained SoftMatch—occupy the top two positions, with SoftMatch's distribution-fitted shape ahead of CCM's fixed sigmoid at the two smallest budgets. We report this ordering in full: it is convergent evidence for the central finding that the weighting shape is the decisive axis, and it delimits what CCM's simpler fixed-sharpness design does and does not buy. We also describe the settings where curriculum-based methods as a group do not lead (SVHN). CCM adds no inference-time cost (Section 5.7).
2. We report a negative result. A natural way to combat confirmation bias is to trust only pseudo-labels that are stable under perturbation—that is, pseudo-labels on which

two independent strong augmentations agree. We implement this idea in two forms: as an admission unlock (agreement admits a sample at a lower threshold) and as reliability reweighting (confident-but-unstable pseudo-labels are down-weighted). In our setting, neither variant improves accuracy, and reweighting consistently reduces it. Pseudo-label diagnostics show that augmentation agreement is a positive correctness signal but, at extreme scarcity, its absolute discriminative level appears too low to filter on safely. We frame this as a limitation of stability-based pseudo-label selection in this specific regime rather than as a universal claim.

3. Through cumulative ablation and a three-dataset study (CIFAR-10, SVHN, and CIFAR-100), we characterize when each ingredient helps in our protocol: soft weighting is decisive only at the smallest budgets and reverses sign as labels grow; curriculum thresholding helps strongly on CIFAR-10, hurts on the easier SVHN, and fails outright on CIFAR-100 at 2500 labels, where its per-class statistics become too thin to be informative; dual-view consistency gives small, broadly consistent gains. We view these ingredient-level observations as practical guidance rather than as general theorems.

We deliberately adopt a reduced compute budget rather than the multi-hundred-GPU-hour schedules of the original SSL papers. This makes every result reproducible on a single commodity GPU in minutes and reflects a second, often-ignored data-related challenge—practitioners with limited compute cannot afford million-step training—while preserving a fair, identical-budget comparison across all methods. We state explicitly that the resulting absolute accuracies are lower than fully converged published numbers; our claims concern relative behavior under a shared budget.

The remainder of this paper is organized as follows. Section 2 reviews related work, including recent developments from 2024 to 2026. Section 3 presents CCM, the unified weighting view, and the cross-view-agreement variants. Section 4 describes the experimental protocol, including the full hardware and software environment. Section 5 reports the main comparison, statistical tests, the ablation, the negative result and its diagnostics, calibration, cross-dataset generalization, and computational cost. Section 6 discusses the findings and their limitations, and Section 7 concludes.

2. Related Work

The idea that a classifier’s predictions should be invariant to input perturbations is the common root of modern consistency-based SSL. The Pi-model and Temporal Ensembling [10,11] penalize disagreement between differently perturbed forward passes of the same image; Mean Teacher [12] stabilizes the target by averaging model weights into an exponential-moving-average (EMA) teacher; and UDA [13] shows that the strength of the perturbation matters, pairing consistency with aggressive learned augmentation. We reuse the EMA teacher from this line as the evaluation model for every method in our study.

Consistency alone, however, does not decide which unlabeled examples deserve a training signal. That question drove the development of pseudo-labeling with confidence thresholds. Self-training on the model’s own hard predictions is one of the oldest ideas in semi-supervised learning [3], and MixMatch [14] combined it with augmentation and MixUp into a practical recipe. FixMatch [5] distilled the recipe to its now-standard form—a fixed high-confidence threshold gating a weak-to-strong consistency loss—and its simplicity made it the reference point for a family of successors that differ chiefly in how the threshold is set. FlexMatch [6] replaces the fixed threshold with per-class curriculum thresholds that track each class’s learning status; FreeMatch [7] makes the threshold self-adaptive; and SoftMatch [8] abandons the hard cutoff altogether in favor of a continuous confidence weight drawn from a truncated Gaussian, trading pseudo-label quantity against quality.

Section 3.4 places these designs, and CCM, in a single weighting framework and states precisely where CCM differs.

This line has continued to develop in the most recent literature. Zhou et al. [15] grew the admission threshold over training and added pseudo-label dropout to curb error accumulation in medical image classification; Li et al. [16] targeted the barely supervised regime directly, augmenting pseudo-labels when only a handful of annotations per class exist; and Jing et al. [17] selected pseudo-labels under a class-balance constraint for few-shot settings, arguing—as we do—that raw confidence is an unreliable selector when labels are scarce. Hybrid designs also combine consistency with contrastive objectives [18] or with dual training paths [19], and calibration-aware pseudo-label selection has been explored for segmentation [20]. These works refine individual mechanisms, typically under full published training schedules; none of them, to our knowledge, compares the underlying ingredients under a single identical modest budget with paired statistical tests, which is the gap this paper addresses.

Our negative result connects to the analysis of confirmation bias by Arazo et al. [9], who show that pseudo-labeling can reinforce its own errors and mitigate this with mixup and label noise. Our diagnostics sharpen the picture at extreme scarcity: the model’s errors are largely stable across augmentations, so perturbation stability—a signal one would expect to flag unreliable pseudo-labels—separates correct from incorrect ones only weakly. Related is the calibration literature: modern networks are systematically overconfident [21], and confirmation bias compounds the problem. We report expected calibration error alongside accuracy and find that all confidence-filtered methods are far better calibrated than threshold-free self-training.

Finally, our protocol follows the tradition of realistic and unified SSL evaluation. Oliver et al. [22] cataloged how evaluation choices inflate SSL results, and the USB benchmark [23] standardized datasets (CIFAR-10/100, SVHN, STL-10) and backbones (Wide ResNet [24]). We follow these conventions at a reduced training budget, in the same spirit of realistic evaluation. For a broader view of the field, we refer the reader to the surveys cited in Section 1 [3,4].

3. Method

3.1. Problem Setup and Notation

We are given a small labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^N$ and a large unlabeled pool $\mathcal{D}_u = \{u_j\}$ drawn from the same C -class distribution, with N as small as 4 labels per class. Let f_θ be a Wide ResNet-28-2 [24] classifier producing logits and $p_\theta(\cdot | x) = \text{softmax}(f_\theta(x))$. We maintain an exponential-moving-average (EMA) teacher $f_{\theta'}$ with $\theta' \leftarrow m\theta' + (1 - m)\theta$, used as the evaluation model for every method (a standard, stabilized test-time average). We write $\mathcal{A}_w$ for a weak augmentation (random crop and, for CIFAR-10/100, horizontal flip) and $\mathcal{A}_s$ for a strong augmentation (weak augmentation followed by RandAugment [25] and Cutout).

3.2. Background: FixMatch

For an unlabeled batch of size μB , FixMatch forms a pseudo-label from the weak view (Equation (1)) and trains the strong view against it when the confidence exceeds a fixed threshold τ (e.g., 0.95); the corresponding unsupervised loss is given by Equation (2).

$$q_j = \arg \max_c p_\theta(c | \mathcal{A}_w(u_j)), \quad \hat{p}_j = \max_c p_\theta(c | \mathcal{A}_w(u_j)). \quad (1)$$

$$\mathcal{L}_u^{\text{FixMatch}} = \frac{1}{\mu B} \sum_j \mathbf{1}[\hat{p}_j \geq \tau] H(q_j, p_\theta(\cdot | \mathcal{A}_s(u_j))), \quad (2)$$

where H is cross-entropy. The total loss is $\mathcal{L} = \mathcal{L}_x + \lambda_u \mathcal{L}_u$ with $\mathcal{L}_x$ the supervised cross-entropy on weakly augmented labeled data.

3.3. CCM

CCM keeps the weak/strong consistency structure and changes which pseudo-labels are trusted and by how much. The pseudo-label is the student's weak-view prediction q_j with confidence $\hat{p}_j$ as above.

- (a) Per-class curriculum threshold. Following FlexMatch [6], we maintain a running estimate σ_c of how many unlabeled examples are confidently predicted as class c (an EMA over batches of the confident-prediction histogram). The per-class threshold is given by Equation (3), so that slowly learned classes (small σ_c) face a lower admission bar.

$$\beta_c = \frac{\sigma_c}{\max_{c'} \sigma_{c'}}, \quad \tau_c = \frac{\beta_c}{2 - \beta_c} \tau. \quad (3)$$

- (b) Smooth confidence weighting. Rather than the hard mask $\mathbf{1}[\hat{p}_j \geq \tau_{q_j}]$, each sample contributes a smooth weight defined in Equation (4), with sharpness parameter γ . Throughout the rest of this paper, $\sigma(\cdot)$ denotes the logistic sigmoid function and should not be confused with the running per-class count σ_c introduced in (a) above (which is always written with a class subscript).

$$w_j = \sigma(\gamma(\hat{p}_j - \tau_{q_j})). \quad (4)$$

Borderline pseudo-labels just below the curriculum threshold are included with a small weight instead of being discarded outright—the key mechanism behind CCM's gains in the extreme-scarcity regime, where few samples pass any hard threshold.

- (c) Dual-view consistency. We draw two independent strong augmentations $\mathcal{A}_s^{(1)}, \mathcal{A}_s^{(2)}$ and average the consistency loss over both (Equation (5)), which reduces the variance of the pseudo-label gradient.

$$\mathcal{L}_u^{\text{CCM}} = \frac{1}{\mu B} \sum_j w_j \frac{1}{2} \left(H(q_j, p_{\theta}(\cdot | \mathcal{A}_s^{(1)}(u_j))) + H(q_j, p_{\theta}(\cdot | \mathcal{A}_s^{(2)}(u_j))) \right). \quad (5)$$

All hyperparameters ($\tau = 0.95$, $\gamma = 10$, $\lambda_u = 1$, $m = 0.999$) are fixed across every dataset, budget, and seed. The second strong view is used only during training; at inference, CCM is an ordinary Wide ResNet, so it adds no deployment cost (quantified in Section 5.7).

Algorithm 1 summarizes one CCM optimization step end-to-end and ties together the equations above with the implementation. Lines 1–2 compute the supervised loss $\mathcal{L}_x$ on the weakly augmented labeled batch. The algorithm uses the implementation symbols `conf_j` and `CE` for what the equations call $\hat{p}_j$ (the weak-view confidence of Equation (1)) and H (cross-entropy); the two are the same quantities, written in code-like form for the pseudocode. Lines 3–9 run entirely under `no_grad` and produce all the per-sample admission information needed for the unsupervised loss: line 4 gives the softmax distribution of the weak-view prediction; line 5 reads off the pseudo-label q_j and its confidence $\hat{p}_j$ (i.e., `conf_j` in the pseudocode); line 6 updates the FlexMatch class counts σ_c with an EMA over batches; lines 7–8 derive the per-class curriculum threshold τ_c of Equation (3); and line 9 computes the smooth confidence weight w_j of Equation (4). Lines 10–11 evaluate the dual-view consistency loss $\mathcal{L}_u^{\text{CCM}}$ of Equation (5) on two independent strong-augmentation views. Lines 12–13 finally apply the SGD student update on $\mathcal{L}_x + \lambda_u \mathcal{L}_u^{\text{CCM}}$ and the EMA teacher update on θ' . The two negative-result variants of Section 5.3 modify only line 9: the admission-unlock variant takes the maximum of w_j and an agreement-gated lower-

threshold weight, and the reliability-reweighting variant multiplies w_j by the agreement-modulated factor $\rho + (1 - \rho) a_j$, with a_j defined in Equation (7).

Algorithm 1 CCM training step.

```

Require: student f_theta, EMA teacher f_theta', labeled batch (X_l, Y_l),
        unlabeled batch U; threshold tau, sharpness gamma, weight lambda_u,
        EMA decay m, learning rate eta; running per-class counts sigma_c.

1: x_w <- WeakAug(X_l)
2: L_x <- CE(f_theta(x_w), Y_l)                                # supervised loss

3: with no_grad:                                              # pseudo-label from weak view
4:   p <- softmax(f_theta(WeakAug(U)))
5:   conf_j, q_j <- max(p_j), argmax(p_j)
6:   sigma_c <- 0.99 * sigma_c + 0.01 * hist(q_j | conf_j >= tau)
7:   beta_c <- sigma_c / max_{c'} sigma_{c'}
8:   tau_c <- (beta_c / (2 - beta_c)) * tau                    # FlexMatch mapping
9:   w_j <- sigmoid(gamma * (conf_j - tau_c[q_j]))             # soft weight

10: s1, s2 <- StrongAug(U), StrongAug(U)                     # two independent strong views
11: L_u <- (1/|U|) * sum_j w_j * 0.5 * ( CE(f_theta(s1_j), q_j)
                                         + CE(f_theta(s2_j), q_j) )

12: theta <- theta - eta * grad_theta (L_x + lambda_u * L_u)  # student update
13: theta' <- m * theta' + (1 - m) * theta                    # EMA teacher update

# Negative-result variants (NOT used by CCM): replace line 9 with one of
# (i) unlock: w_j <- max(w_j, a_j * sigmoid(gamma*(conf_j - kappa*tau_c[q_j])))
# (ii) reweight: w_j <- w_j * (rho + (1 - rho) * a_j)
# with a_j = 1[argmax f(s1_j)==q_j] AND 1[argmax f(s2_j)==q_j].
  
```

3.4. A Unified View: Pseudo-Label Weighting Functions

The methods compared in this paper differ, at their core, in a single design decision: the function g that maps a sample's weak-view confidence $\hat{p}_j$ (and the training state) to a weight on its consistency loss,

$$\mathcal{L}_u = \frac{1}{\mu_B} \sum_j g(\hat{p}_j; \text{state}) \ell_j, \quad (6)$$

where ℓ_j is the strong-view cross-entropy toward the pseudo-label. FixMatch instantiates g as a fixed step function, $g = \mathbf{1}[\hat{p}_j \geq \tau]$; FlexMatch as a class-conditional step function, $g = \mathbf{1}[\hat{p}_j \geq \tau_{q_j}]$, whose per-class centers track learning progress; SoftMatch as a truncated-Gaussian weight whose center and width follow global running statistics of the confidence distribution; and CCM as a logistic weight with a class-conditional center, $g = \sigma(\gamma(\hat{p}_j - \tau_{q_j}))$. Table 1 summarizes the design space.

Table 1. Pseudo-label weighting designs as instances of the unified weighting function of Equation (6). CCM combines a class-conditional center (from FlexMatch) with a smooth shape (in the spirit of SoftMatch, but simpler) and adds a second strong view.

Method	Threshold/Center	Weighting Shape	Per-Class State	Strong Views
FixMatch	fixed τ	step (hard)	none	1
FlexMatch	curriculum τ_c	step (hard)	class counts σ_c	1
FreeMatch	self-adaptive	step (hard, adaptive)	class statistics	1
SoftMatch	global mean μ_t	truncated Gaussian	none (global)	1
CCM	curriculum τ_c	logistic (smooth)	class counts σ_c	2

This view makes the division of labor among CCM's three ingredients explicit and explains why they are complementary rather than redundant. The curriculum threshold

moves the center of the weighting function per class: a class that has produced few confident predictions receives a lower center, so its samples are not starved of training signal. The smooth logistic shape supplies a gradient around that center: samples just below the threshold contribute proportionally rather than being discarded, which matters precisely when almost no sample clears any hard bar. The dual strong view acts on a different axis entirely—not on which samples are weighted but on the noise of each sample’s loss: averaging two independent strong-view losses halves the variance of the per-sample gradient estimate, which is most valuable exactly when soft weighting deliberately admits lower-confidence (noisier) targets. In short, the center decides which classes get signals, the shape decides how many signals borderline samples get, and the second view controls how noisy that signal is. The ablation in Section 5.2 tests each of these axes separately and confirms that their contributions are distinct and budget-dependent.

Two differences from SoftMatch are worth stating precisely. First, CCM’s center is class-conditional where SoftMatch’s is global: under extreme scarcity, classes learn at visibly different rates (Section 5.6), and a global center cannot compensate a slow class without also admitting noise in fast classes. Second, CCM’s shape has a single fixed sharpness γ and no running mean or variance estimation, making it a strict plug-in on top of an existing FlexMatch implementation; we deliberately avoid the extra estimation machinery so that the ablation isolates the qualitative effect of soft inclusion rather than the quality of a fitted Gaussian. Comparing CCM against the published FreeMatch/SoftMatch numbers would require matching their training schedules (millions of steps with per-method-tuned hyperparameters); under our shared 4000-step budget, all methods are far from their fully trained ceilings, which is why we re-train every compared method—including SoftMatch itself (Section 5.1)—under the identical protocol.

3.5. A Natural Idea That Does Not Work: Cross-View Agreement

Confirmation bias suggests trusting only pseudo-labels that are stable under perturbation. We define the agreement indicator a_j as in Equation (7) and use it two ways.

$$a_j = \mathbf{1}[\arg \max_c p_\theta(c \mid \mathcal{A}_s^{(1)}(u_j)) = q_j] \cdot \mathbf{1}[\arg \max_c p_\theta(c \mid \mathcal{A}_s^{(2)}(u_j)) = q_j]. \quad (7)$$

- (i) Admission unlock: agreeing samples are admitted at a lowered threshold $\kappa\tau_{q_j}$ (with $\kappa = 0.6$ in all our runs), so the per-sample weight becomes $\max(w_j, a_j \sigma(\gamma(\hat{p}_j - \kappa\tau_{q_j})))$, which only adds pseudo-labels.
- (ii) Reliability reweighting: confident-but-unstable pseudo-labels are down-weighted by the factor $w_j \cdot (\rho + (1 - \rho)a_j)$ with $\rho = 0.5$ in all our runs, which only reweights within the admitted set.

These two hyperparameters (κ and ρ) are introduced for the variants of this section only; CCM itself does not use them. In the unified view of Section 3.4, both variants modulate the weighting function g by the stability signal a_j ; the reason this modulation fails empirically is quantified in Sections 5.3 and 5.4. We evaluate both variants in Section 5.3; neither improves over CCM in our setting, and the reweighting variant consistently hurts—Section 6 develops a plausible explanation.

4. Experimental Protocol

Our protocol is designed around one principle: every method sees the identical data; backbone; augmentation; optimizer; and most importantly, the identical compute budget, so that differences in accuracy are attributable to the methods’ mechanisms rather than to their tuning or schedule.

Datasets. The primary benchmark is CIFAR-10 [26]; SVHN [27] and CIFAR-100 [26] serve as generalization checks on an easier and a harder dataset, respectively. CIFAR-100's hundred classes are a deliberate stress test for the per-class curriculum mechanism, which must then track ten times more classes from the same number of unlabeled batches. For each label budget, we draw a class-balanced labeled subset with a seed-specific random generator; the entire training set serves as the unlabeled pool, following the FixMatch protocol. Because the identity of the labeled examples is itself part of the difficulty under label scarcity, the variance across seeds (which draw different labeled subsets) is reported as part of every result.

Backbone and optimization. All methods share a Wide ResNet-28-2 backbone (1.47 M parameters), trained with SGD with Nesterov momentum 0.9, weight decay 5×10^{-4} , base learning rate 0.03 under a cosine schedule, labeled batch size 64, unlabeled ratio $\mu = 7$, and automatic mixed precision. Weak/strong augmentation and the EMA teacher follow Section 3.

Environment. All experiments were run on a single NVIDIA GeForce RTX 4090 (24 GB) GPU using Python 3.8, PyTorch 2.4.1 with CUDA 12.1, torchvision 0.19.1, NumPy 1.24.4, and SciPy 1.10.1. Under this environment, each 4000-step run completes in one to five minutes depending on the method (see Section 5.7), so the full study is reproducible on one commodity GPU in a few hours.

Compute-constrained budget. Every run is trained for 4000 optimization steps—orders of magnitude fewer than the original SSL papers—and evaluated with the EMA teacher. We report the median of the last five evaluation checkpoints, which avoids selecting on the test set. All methods use the identical budget, so comparisons are fair; absolute accuracies are correspondingly lower than fully converged literature numbers, by design.

Methods compared. Supervised-only, threshold-free pseudo-label self-training [3], Mean Teacher [12], FixMatch [5], FlexMatch [6], SoftMatch [8] (re-implemented in its default uniform-prior form and re-trained under our budget), and CCM. Every method shares the backbone, optimizer, augmentations, and budget.

Ablation. Starting from FixMatch, we add components cumulatively: (i) FixMatch (base) $\rightarrow$ (ii) + curriculum threshold (=FlexMatch) $\rightarrow$ (iii) + second strong view (dual-view consistency) $\rightarrow$ (iv) + smooth confidence weighting (=CCM). We then separately add each of the two cross-view-agreement mechanisms of Section 3.5 on top of CCM.

Metrics. Test accuracy (mean $\pm$ std over three seeds), expected calibration error (ECE, 15 bins) [21], and two confirmation-bias diagnostics: the mask rate (fraction of the unlabeled batch with weight above 0.5) and the pseudo-label accuracy (fraction of those pseudo-labels matching the withheld ground truth—used for analysis only, never for training).

5. Results

5.1. Main Comparison on CIFAR-10

Table 2 reports test accuracy across label budgets. In the extreme-scarcity regime, the two smooth-weighting designs clearly lead the field: with only 40 labels (4 per class), SoftMatch reaches 37.04% and CCM 35.14%, both well ahead of the hard-threshold methods FlexMatch (31.84%) and FixMatch (30.55%), and far above supervised-only (23.76%) and pseudo-label (24.30%). CCM improves on FlexMatch—the method it directly extends—by 3.30 pp ($p = 0.001$), which isolates the value of smooth inclusion on top of an otherwise identical curriculum design; SoftMatch, whose truncated-Gaussian weight is additionally fitted to the running confidence distribution, performs better still (+1.90 pp over CCM, $p = 0.049$). The same ordering holds at 250 labels (SoftMatch 59.52%, FlexMatch 57.21%, CCM 55.56%, and FixMatch 52.81%), while at 4000 labels, the three adaptive methods converge and CCM is numerically ahead (CCM 79.80%, FlexMatch 79.97%, and SoftMatch

78.27%; CCM-vs-SoftMatch $p = 0.063$, not significant). We highlight what this ordering does and does not say. It does say—twice over, with two independent designs—that a smooth weighting shape is the decisive ingredient under extreme scarcity and that fitting the shape to the confidence distribution (SoftMatch) extracts more from it than a fixed sigmoid sharpness (CCM). It does not change the component-level findings of Sections 5.2–5.4, which use CCM precisely because its fixed-sharpness design isolates the qualitative effect of soft inclusion from shape fitting. Two further patterns are worth noting. First, on average, CCM improves over the FixMatch recipe it extends at every budget (+4.59/+2.75/+5.68 pp at 40/250/4000), although not all of these gaps are statistically significant (see the Statistical Significance paragraph below). Second, the simplest methods behave non-monotonically with respect to the budget: at 4000 labels, plain FixMatch (74.12%) sits slightly below supervised-only training (76.86%)—a pattern consistent with the hypothesis that a fixed high-confidence threshold may inject some harmful pseudo-labels once enough labels are available, a behavior the adaptive methods do not exhibit in our runs. The accuracy–budget trend is visualized in Figure 1.

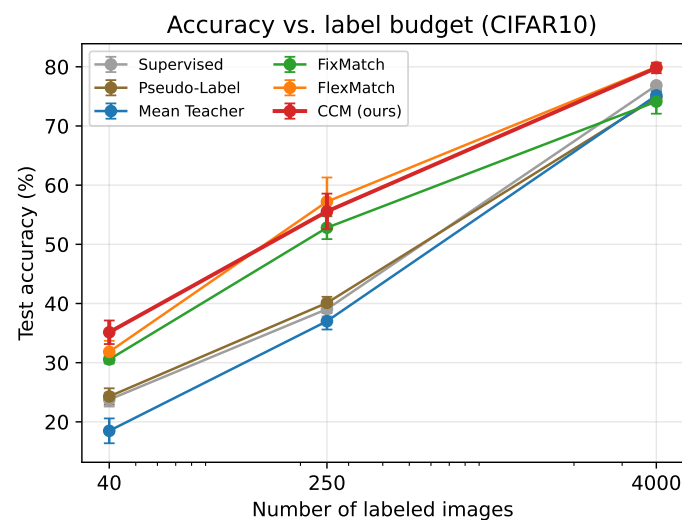


Figure 1. Test accuracy vs. label budget on CIFAR-10 (mean over three seeds). CCM leads in the extreme-scarcity regime and matches FlexMatch as labels grow.

Table 2. Test accuracy (%) on CIFAR-10 across label budgets; median of the last five evaluation checkpoints, mean ± std over three seeds.

Method	40 Labels	250 Labels	4000 Labels
Supervised	23.76 ± 1.17	39.05 ± 1.23	76.86 ± 0.40
Pseudo-Label	24.30 ± 1.37	40.10 ± 1.03	74.77 ± 1.01
Mean Teacher	18.48 ± 2.10	37.01 ± 1.40	75.16 ± 0.67
FixMatch	30.55 ± 0.68	52.81 ± 1.93	74.12 ± 2.05
FlexMatch	31.84 ± 1.86	57.21 ± 4.08	79.97 ± 0.19
SoftMatch	37.04 ± 2.58	59.52 ± 3.00	78.27 ± 0.69
CCM (ours)	35.14 ± 1.99	55.56 ± 3.01	79.80 ± 0.88

Statistical significance. With three seeds per cell—itself a data-related challenge (see Section 6 on label-subset variance)—statistical power is limited, so we report paired tests on the same per-seed labeled subsets rather than two-sample tests. Table 3 gives the 95% Welch t-confidence interval for each method’s mean accuracy and the paired t -test p -value of CCM versus each baseline.

Table 3. Statistical comparison of CCM against each baseline on CIFAR-10 (per-budget 95% Welch *t*-confidence interval; paired *t*-test and Wilcoxon signed-rank *p*-values over three seeds with matched labeled subsets).

Budget	Method	Mean	95% CI (Welch)	vs. CCM Δ	Paired <i>t p</i>	Wilcoxon <i>p</i>
40	Supervised	23.76	[20.20, 27.32]	+11.38	0.015	0.109
40	Pseudo-Label	24.30	[20.12, 28.47]	+10.84	0.021	0.109
40	Mean Teacher	18.48	[12.10, 24.87]	+16.66	0.010	0.109
40	FixMatch	30.55	[28.47, 32.62]	+4.59	0.105	0.109
40	FlexMatch	31.84	[26.18, 37.49]	+3.30	0.001	0.109
40	SoftMatch	37.04	[29.18, 44.89]	−1.90	0.049	0.109
40	CCM (ours)	35.14	[29.08, 41.20]	—	—	—
250	Supervised	39.05	[35.31, 42.79]	+16.51	0.020	0.109
250	Pseudo-Label	40.10	[36.98, 43.23]	+15.46	0.015	0.109
250	Mean Teacher	37.01	[32.74, 41.28]	+18.55	0.019	0.109
250	FixMatch	52.81	[46.94, 58.68]	+2.75	0.476	0.593
250	FlexMatch	57.21	[44.80, 69.63]	−1.65	0.164	0.109
250	SoftMatch	59.52	[50.40, 68.64]	−3.96	0.002	0.109
250	CCM (ours)	55.56	[46.41, 64.71]	—	—	—
4000	Supervised	76.86	[75.63, 78.08]	+2.94	0.060	0.109
4000	Pseudo-Label	74.77	[71.70, 77.84]	+5.03	0.033	0.109
4000	Mean Teacher	75.16	[73.11, 77.21]	+4.64	0.049	0.109
4000	FixMatch	74.12	[67.89, 80.35]	+5.68	0.061	0.109
4000	FlexMatch	79.97	[79.38, 80.55]	−0.17	0.845	1.000
4000	SoftMatch	78.27	[76.16, 80.38]	+1.53	0.063	0.109
4000	CCM (ours)	79.80	[77.11, 82.49]	—	—	—

The headline finding—CCM > FlexMatch at 40 labels—is significant at $p = 0.001$. The CCM > FixMatch gain at 40 labels (+4.59 pp) does not reach $p < 0.05$ ($p = 0.105$), and the CCM vs. FlexMatch result at 250 labels (−1.65 pp) and 4000 labels (−0.17 pp) are clearly non-significant ($p = 0.164$ and $p = 0.845$), confirming the “tied at moderate budgets” reading. The SoftMatch comparison is reported with the same rigor: SoftMatch is significantly ahead of CCM at 40 labels (+1.90 pp, $p = 0.049$) and at 250 labels (+3.96 pp, $p = 0.002$), and statistically tied with CCM at 4000 labels (−1.53 pp in CCM’s favor, $p = 0.063$); it is also significantly ahead of FlexMatch at 40 labels (+5.20 pp, $p = 0.010$). We also report the Wilcoxon signed-rank *p*-value, which by construction cannot reach $p < 0.05$ with $n = 3$ (the smallest achievable Wilcoxon *p* for $n = 3$ is 0.109), so it is included only as a robustness pointer rather than a power-bearing test.

5.2. Component Ablation

Table 4 traces the cumulative effect of each component on CIFAR-10 (Figure 2). Read through the unified view of Section 3.4, the ablation separates the three axes of the weighting function. Moving the center per class (curriculum thresholding) is the largest single contributor at 250 labels (+4.40 pp over FixMatch) and a moderate one at 40 (+1.29 pp). Reducing per-sample gradient noise (the second strong view) gives a small but consistent gain at both budgets (+0.21/+0.89 pp). Changing the shape from a step to a smooth logistic is decisive under extreme scarcity: at 40 labels, it adds +3.09 pp—the single largest step there—because softly admitting borderline pseudo-labels recovers supervision that a hard curriculum threshold discards when almost nothing is classified confidently. At 250 labels, however, the same soft inclusion admits more label noise than it is worth and costs −2.54 pp relative to the hard-threshold variant. Soft weighting is thus a budget-dependent quantity–quality trade-off that pays off precisely in the regime CCM targets; we keep it as a fixed part of CCM and report this trade-off transparently rather than tuning it per budget.

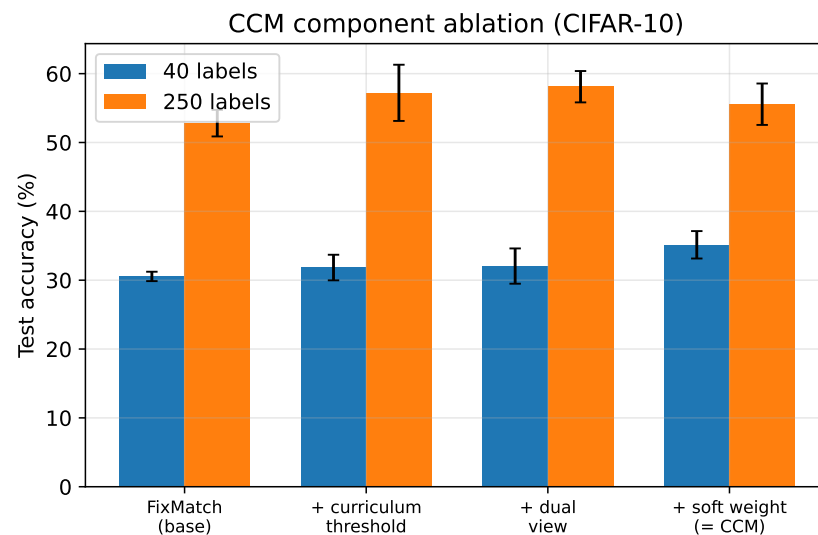


Figure 2. Cumulative component ablation of CCM on CIFAR-10. Soft weighting is decisive at 40 labels; at 250 labels, it trades quantity for quality.

Table 4. Cumulative component ablation of CCM on CIFAR-10 (test accuracy%, mean $\pm$ std over three seeds).

Variant	40 Labels	250 Labels
FixMatch (base)	30.55 $\pm$ 0.68	52.81 $\pm$ 1.93
+curriculum threshold	31.84 $\pm$ 1.86	57.21 $\pm$ 4.08
+dual view	32.05 $\pm$ 2.56	58.10 $\pm$ 2.28
+soft weighting (=CCM)	35.14 $\pm$ 1.99	55.56 $\pm$ 3.01

5.3. The Cross-View-Agreement Negative Result

A natural way to combat confirmation bias is to trust only pseudo-labels that are stable under perturbation. We added the two cross-view-agreement mechanisms of Section 3.5 on top of CCM, with both formulations evaluated on the same seeds and budget so the comparison is paired (Figure 3). Table 5 summarizes the three configurations side by side; neither helps, and both range from neutral to clearly hurting:

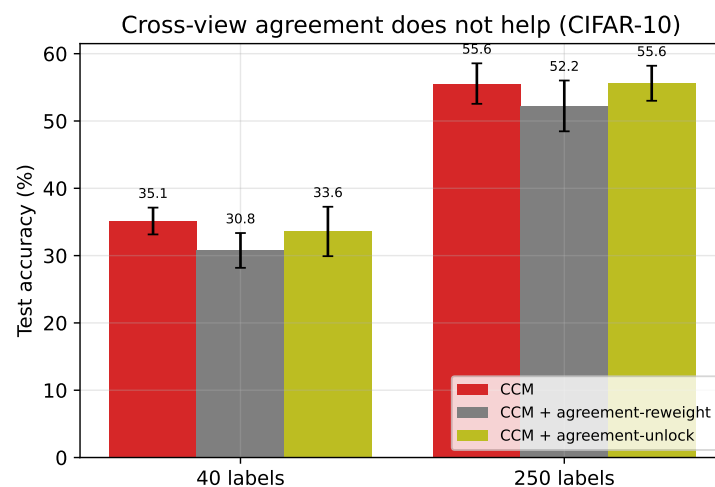


Figure 3. Negative result: adding cross-view agreement on top of CCM (reweighting and admission-unlock formulations) does not improve and tends to hurt accuracy at every budget.

Table 5. Cross-view-agreement variants added on top of CCM (CIFAR-10, mean $\pm$ std over three seeds). Neither reweighting nor admission unlock improves over CCM.

Variant	40 Labels	250 Labels	4000 Labels
CCM	35.14 $\pm$ 1.99	55.56 $\pm$ 3.01	79.80 $\pm$ 0.88
CCM + agreement (reweight)	30.77 $\pm$ 2.58	52.24 $\pm$ 3.77	76.26 $\pm$ 1.00
CCM + agreement (unlock)	33.59 $\pm$ 3.67	55.61 $\pm$ 2.60	—

- The reliability-reweighting variant—which down-weights confident-but-cross-view-unstable pseudo-labels—reduces accuracy at every budget: by -4.37 pp at 40 labels (CCM 35.14 $\rightarrow$ 30.77), -3.32 pp at 250, and -3.54 pp at 4000.
- The admission-unlock variant—which admits agreeing low-confidence pseudo-labels at a lowered threshold $\kappa\tau_c$ —is essentially neutral: -1.55 pp at 40 labels (35.14 $\rightarrow$ 33.59) and $+0.05$ pp at 250. It never improves over CCM in our runs.

The diagnostic that explains this is the agreement-conditional pseudo-label accuracy, which we measure directly during CCM training (Section 5.4 reports it for this section): among the unlabeled samples used by CCM at 40 labels, agreeing pseudo-labels are more accurate than disagreeing ones ($\approx 43.9\%$ vs. 29.4%), confirming that cross-view stability is a real reliability signal. The trouble is the absolute level: agreement raises pseudo-label accuracy to roughly half, which is still too noisy to safely up-weight (reweighting) or to safely admit at a lower threshold (unlock). Conversely, the 29.4% accuracy of the disagreeing pseudo-labels is still informative on average—discarding or down-weighting them removes useful gradient. Section 6 develops this argument.

In short, in our setting, the failure is not that the agreement signal is useless (it is mildly informative), but that its absolute discriminative level appears too low—at this label scarcity and under this compute budget—to justify either filtering or coverage expansion. We frame this as a limit of stability-based pseudo-label selection in this specific regime rather than as a universal claim about the idea; agreement-based filtering may well be useful in less label-scarce or longer-trained settings.

5.4. Pseudo-Label Quality and Confirmation Bias

Figure 4 tracks two diagnostics over training: the mask rate (fraction of the unlabeled batch used) and the pseudo-label accuracy (fraction of used pseudo-labels matching the withheld ground truth). At 40 labels, FixMatch’s fixed threshold admits a minority of the batch (mask rate $\approx 43\%$) at a pseudo-label accuracy of $\approx 40\%$, whereas the curriculum methods admit far more—FlexMatch and CCM reach $\approx 74\%$ —at a similar or slightly lower pseudo-label accuracy (36.8% for FlexMatch, 40.7% for CCM). Two observations follow. First, roughly 60% of the pseudo-labels CCM trains on at 40 labels are wrong, yet they are collectively informative: CCM still reaches 35% test accuracy, far above the 24% of supervised-only training. Second, soft weighting lets CCM match FlexMatch’s high usage while raising pseudo-label accuracy slightly, which is the mechanism behind its gain. At 250 labels, the same picture holds at higher quality (FixMatch 75.8% pseudo-label accuracy at 55.8% usage; CCM 65.3% at 87.9% usage).

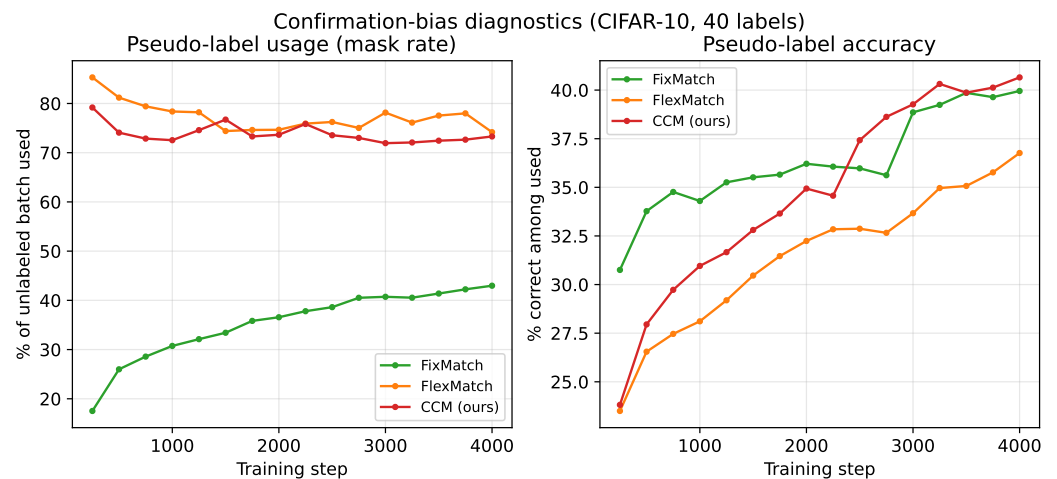


Figure 4. Confirmation-bias diagnostics at 40 labels: pseudo-label usage (**left**) and pseudo-label accuracy among used samples (**right**).

Precision–coverage analysis. Figure 5 sweeps the confidence threshold over the final model’s weak-view predictions on the unlabeled pool and plots pseudo-label precision against coverage for FixMatch, FlexMatch, and CCM at 40 labels. The curves make the quantity–quality trade-off of Section 5.2 visible directly and add a finding: the CCM-trained model offers strictly higher pseudo-label precision than the FlexMatch- and FixMatch-trained models at every coverage level, indicating that soft training improves the pseudo-label generator itself rather than merely reweighting its output. Yet, even the best curve stays below 50% precision at any meaningful coverage in this regime—the empirical reason why hard filtering (whether by a fixed threshold or by cross-view agreement) cannot rescue pseudo-label quality at extreme scarcity. Figure 6 complements this with the confidence and weight distributions at the start, middle, and end of training: the confidence distribution is strongly bimodal, and CCM’s logistic weighting maps the mid-confidence mass to small-but-nonzero weights rather than discarding it.

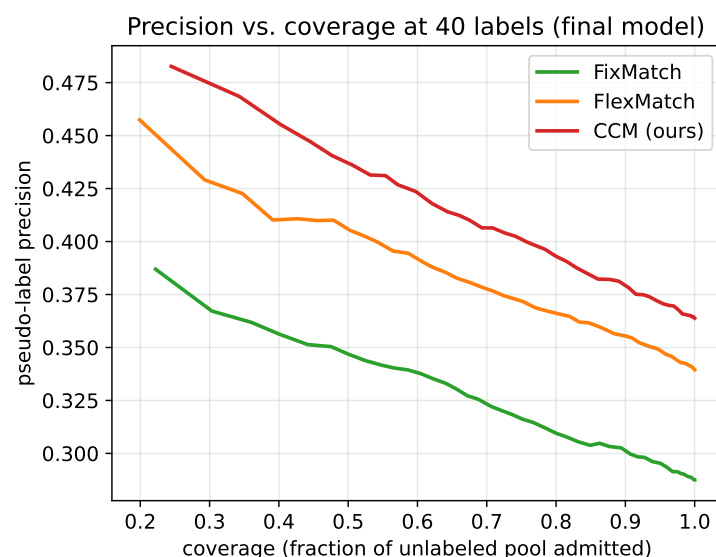


Figure 5. Pseudo-label precision vs. coverage at 40 labels as the confidence threshold is swept over the final model’s weak-view predictions. The CCM-trained model dominates at every coverage level, yet no operating point offers high precision at meaningful coverage in this regime.

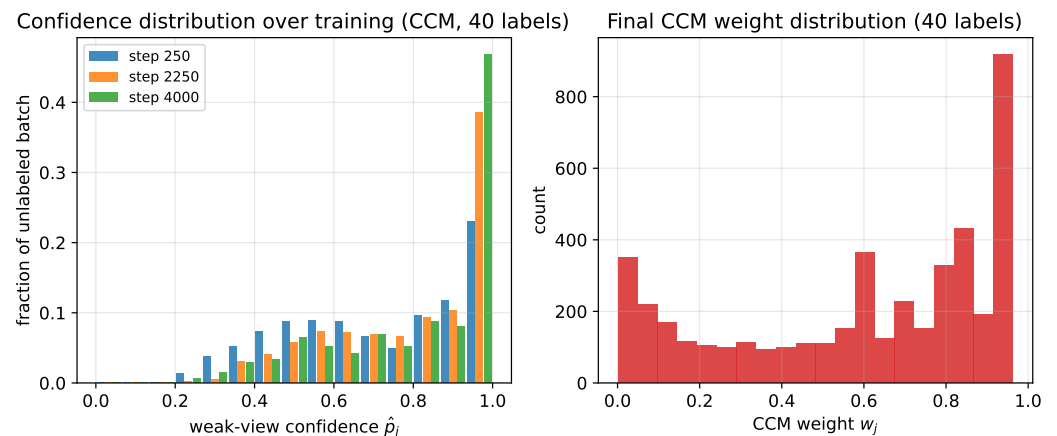


Figure 6. Weak-view confidence distributions at the start, middle, and end of CCM training at 40 labels (**left**) and the resulting distribution of CCM’s logistic weights at the end of training (**right**).

Agreement-conditional pseudo-label accuracy. As a direct test of the “stable = correct” hypothesis underlying Section 5.3, we additionally instrumented CCM training to log, per batch, the agreement indicator and the agreement-conditional pseudo-label accuracy on the (held-out, diagnostics-only) true labels (Figure 7). Among the unlabeled samples CCM actually uses at 40 labels, 55.4% show cross-view agreement; agreeing pseudo-labels are correct 43.9% of the time, while disagreeing ones are correct only 29.4%. Agreement is therefore a real but weak reliability signal: it discriminates correct from incorrect labels, but the absolute correctness of the agreeing set is still close to 50%. This number is the empirical basis for the negative result of Section 5.3.

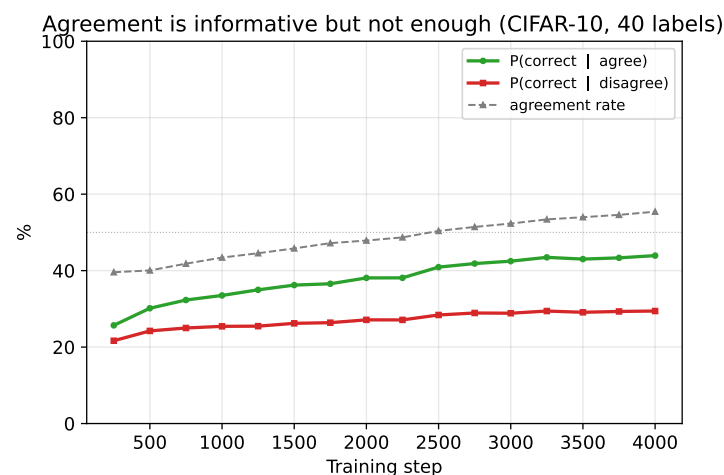


Figure 7. Agreement-conditional pseudo-label accuracy under CCM training at 40 labels. Agreement raises pseudo-label accuracy but the absolute level stays close to 50%.

Qualitative complement. Figure 8 visualizes the penultimate-layer features of the supervised-only baseline (left) and CCM (right) at 250 labels via t-SNE on the test set. Supervised training produces a partially mixed embedding in which several classes overlap; CCM, using the same 250 labels but exploiting the unlabeled pool, recovers a substantially better-separated embedding with most CIFAR-10 classes forming distinct clusters—a visual confirmation that the test-accuracy gain comes from genuinely better representations rather than a re-calibration of the decision layer alone. Figure 9 shows the row-normalized confusion matrix of CCM at the same budget. Errors concentrate on the visually similar cat/dog and automobile/truck pairs—a pattern long documented for CIFAR-10 classifiers—rather than collapsing onto any one class, which would be the calibration

signature of an unchecked confirmation-bias loop. The matrix therefore confirms that the soft-weighting mechanism (Section 5.4) admits noisy pseudo-labels without inducing a systematic class-collapse.

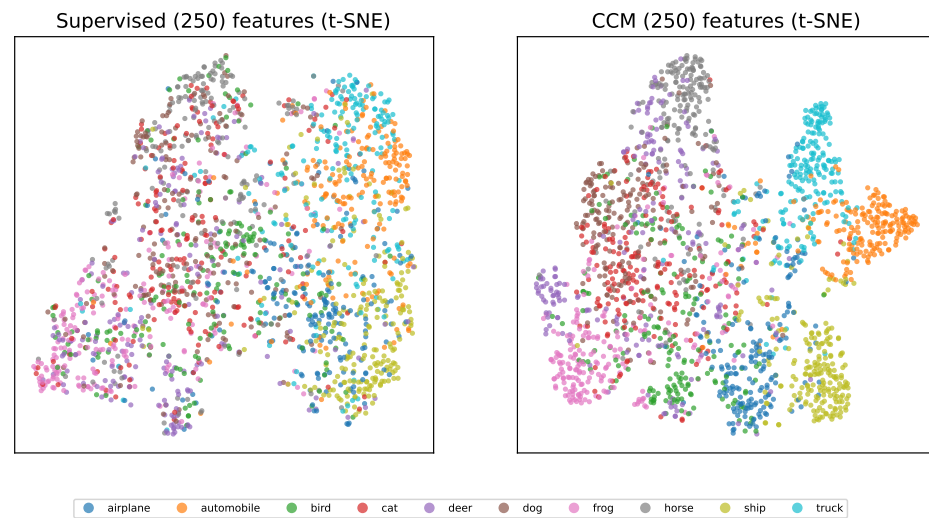


Figure 8. t-SNE of penultimate features at 250 labels: supervised-only (left) vs. CCM (right).

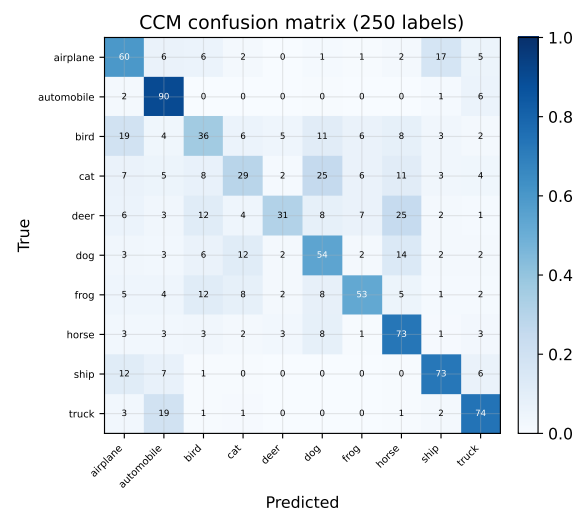


Figure 9. CCM confusion matrix on CIFAR-10 at 250 labels (row-normalized, %).

5.5. Calibration

We measure expected calibration error (ECE, 15 bins) of the final classifier (Figure 10). All consistency methods are far better calibrated than threshold-free self-training: at 40 labels, pseudo-label reaches an ECE of 67% (severely overconfident—the calibration signature of confirmation bias) and supervised-only 52%, whereas FixMatch/FlexMatch/CCM sit at 48/47/45%. CCM attains the lowest ECE at 40 and at 4000 labels (45% and 9%) and is within 3 pp of the best at 250, so its accuracy gains do not come at a calibration cost. The pattern reinforces the confirmation-bias reading: methods that filter or weight pseudo-labels by confidence are markedly less overconfident than naive self-training.

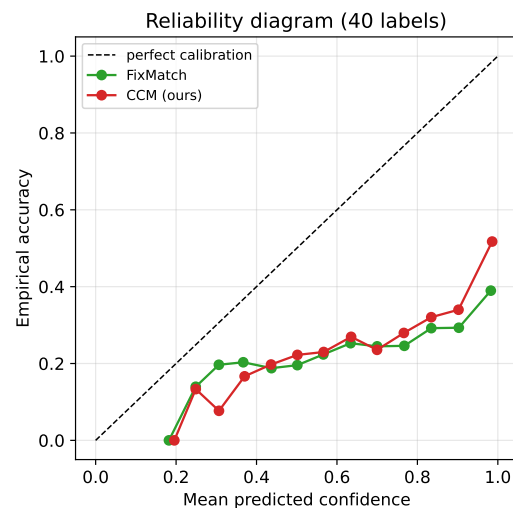


Figure 10. Reliability diagram at 40 labels.

5.6. Generalization Across Datasets

SVHN. On SVHN at 250 labels the ranking differs sharply (Table 6): FixMatch is best (91.82%), while the curriculum-based methods FlexMatch (84.37%) and CCM (84.99%) trail it, even though all SSL methods dwarf supervised-only training (45.08%). SVHN is an easy SSL dataset on which a fixed high threshold already yields very reliable pseudo-labels; the per-class curriculum that helps on CIFAR-10 instead lowers thresholds unnecessarily and admits noise. Figure 11 visualizes the mechanism: on CIFAR-10, at 40 labels, the per-class thresholds start low, climb throughout training, and remain widely spread—the curriculum continuously re-allocates admission bars as slow classes catch up. On SVHN, the trajectories freeze within the first few hundred steps at widely separated levels, most of them well below the fixed threshold: because a few easy digit classes dominate the confident-prediction counts, the remaining classes are held at permanently lowered bars, admitting borderline pseudo-labels that a fixed threshold would reject—on a dataset where high-confidence pseudo-labels are already plentiful. CCM slightly improves on its FlexMatch base (+0.62 pp) but inherits the curriculum’s disadvantage here. We report this as an honest boundary of the method: CCM’s ingredients are tuned for the hard, label-scarce regime and are not universally optimal—choosing SSL components to match the data regime is itself one of the data-related challenges this paper highlights.

CIFAR-100. Table 7 extends the comparison to CIFAR-100 under the identical protocol, at 400 and 2500 labels (4 and 25 per class). At 400 labels, the four methods sit in a narrow band at very low absolute accuracy (supervised 9.21%, FixMatch 9.96%, FlexMatch 10.83%, and CCM 11.01%): under a 4000-step budget the 100-class problem is barely past initialization for every method, and the curriculum-based methods hold a small numerical lead. At 2500 labels, the picture changes sharply and instructively: supervised training (28.94%) and FixMatch (27.54%) behave normally, while the two curriculum-based methods collapse—FlexMatch to 4.23% and CCM to 7.91%. The mechanism is traceable to the curriculum statistics: with 100 classes, the per-class confident-prediction counts σ_c that drive Equation (3) are estimated from roughly ten times fewer confident predictions per class than on CIFAR-10, so many classes sit near-permanently at very low thresholds, and the resulting flood of low-quality pseudo-labels overwhelms the supervised signal—a curriculum-driven confirmation-bias collapse. CCM’s smooth weighting softens the failure relative to FlexMatch (+3.68 pp) but does not rescue it. We report this as a substantive boundary of the curriculum mechanism itself: it presumes per-class statistics dense enough to be informative, an assumption that CIFAR-100 under a short schedule violates. Together

with SVHN, the three-dataset picture is consistent: the curriculum center helps exactly when classes learn at usefully different rates and its statistics are well estimated (CIFAR-10), is redundant when pseudo-labels are already reliable (SVHN), and is actively harmful when its statistics are too thin to be meaningful (CIFAR-100 at 2500 labels under this budget).

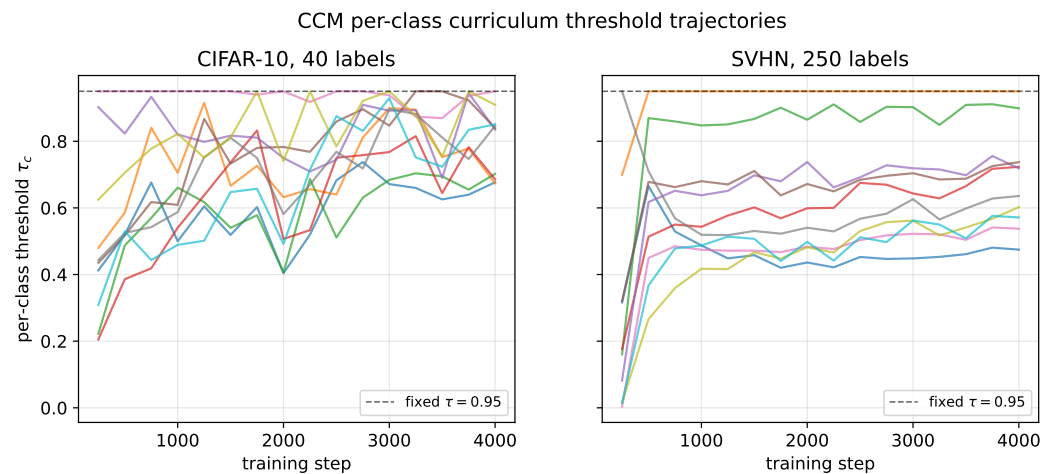


Figure 11. Per-class curriculum threshold trajectories over training for CCM on CIFAR-10 at 40 labels (**left**) and SVHN at 250 labels (**right**). Each colored curve corresponds to one of the ten classes of the respective dataset; the colors serve only to distinguish the classes. On CIFAR-10, the thresholds climb and remain widely spread as the curriculum re-allocates admission bars; on SVHN, they freeze early at levels mostly below the fixed threshold, holding admission bars down on a dataset whose pseudo-labels are already reliable.

Table 6. Generalization to SVHN at 250 labels (mean $\pm$ std over three seeds). FixMatch leads on this easier dataset, illustrating that the curriculum-based methods' advantage depends on dataset difficulty.

Method	SVHN (250 Labels)
Supervised	45.08 $\pm$ 5.92
FixMatch	91.82 $\pm$ 1.14
FlexMatch	84.37 $\pm$ 3.66
CCM (ours)	84.99 $\pm$ 2.59

Table 7. Generalization to CIFAR-100 (test accuracy% and mean $\pm$ std over three seeds) under the same 4000-step protocol, with 4 and 25 labels per class. The 100-class setting stresses the per-class curriculum threshold with ten times fewer confident predictions per class.

Method	400 Labels	2500 Labels
Supervised	9.21 $\pm$ 0.42	28.94 $\pm$ 0.55
FixMatch	9.96 $\pm$ 0.52	27.54 $\pm$ 0.49
FlexMatch	10.83 $\pm$ 0.55	4.23 $\pm$ 0.55
CCM (ours)	11.01 $\pm$ 1.03	7.91 $\pm$ 1.58

5.7. Computational Cost

Table 8 quantifies the claim that CCM adds no inference-time cost. All methods deploy the identical Wide ResNet-28-2 (1,467,610 parameters; 0.429 GFLOPs per 32×32 image at inference) because every CCM ingredient—the curriculum threshold, the logistic weighting, and the second strong view—acts only in the training loss and adds no parameters and no operations to the deployed model. Training cost does differ: the second strong view adds one forward and backward pass over the unlabeled batch, and the measured wall-clock overhead of CCM is a factor of 1.53 over FixMatch under our protocol (69.3 versus 45.2 ms

per step on an RTX 4090). We consider this an acceptable trade-off in the label-scarce regime the method targets, and we report it transparently.

Table 8. Computational cost. Parameters and inference FLOPs are identical across all methods (same deployed backbone); training-time cost is measured via a wall-clock per optimization step on a single RTX 4090 (mean over all runs of each method).

Method	Params (M)	Inference GFLOPs	Training ms/Step	Ratio to FixMatch
Supervised	1.47	0.429	13.1	0.29
Pseudo-Label	1.47	0.429	44.4	0.98
Mean Teacher	1.47	0.429	44.0	0.98
FixMatch	1.47	0.429	45.2	1.00
FlexMatch	1.47	0.429	47.4	1.05
CCM	1.47	0.429	69.3	1.53

6. Discussion

A plausible explanation for the cross-view-agreement failure. The diagnostics of Section 5.4 give a quantitative picture. Cross-view agreement does carry a positive correctness signal—at 40 labels, agreeing pseudo-labels are correct ~43.9% of the time versus ~29.4% for disagreeing ones—but two empirical patterns appear to undermine both interventions of Section 3.5 in our setting. First, the absolute correctness of agreeing pseudo-labels is still roughly one-in-two; up-weighting them (reweighting) or admitting more at a lower threshold (unlock) amplifies a target distribution that remains heavily contaminated, which is exactly what confirmation bias feeds on. Second, the disagreeing pseudo-labels, while less reliable individually, appear to still carry useful average gradient: with three to four labels per class as the only anchor, even ~29.4%-accurate pseudo-labels seem to nudge the decision boundary in roughly the right direction, and removing them via reweighting discards supervision the model can ill afford to lose. The combination—agreement is informative but apparently not informative enough, and disagreement is wrong but not wrong enough—is consistent with the observation that neither formulation improves over CCM and that reweighting consistently hurts. In the language of Section 3.4: modulating the weighting function by a reliability signal only pays off when the signal’s conditional precision clears the noise floor of the admitted pseudo-labels, and at this scarcity, it does not. A tentative generalizable lesson would be that augmentation-invariance is a partial reliability signal whose threshold for being useful as a filter depends on the noise floor of the pseudo-labels themselves, which in turn depends on the label budget; testing this beyond our compute-constrained protocol is left to future work.

When soft weighting helps. Soft confidence weighting is the ingredient that wins the extreme-scarcity regime, and the SoftMatch comparison of Section 5.1 makes the evidence convergent: two independently designed smooth weighting functions—SoftMatch’s distribution-fitted truncated Gaussian and CCM’s fixed-sharpness sigmoid—finish first and second at 40 labels, ahead of every hard-threshold method. Within the unified view of Section 3.4, the additional gap between them (+1.90 pp at 40 labels, $p = 0.049$) measures the value of fitting the shape to the running confidence distribution rather than fixing its sharpness; the price of the fitted shape appears at 4000 labels, where SoftMatch falls numerically behind both curriculum methods. In the CCM ablation, the same soft-inclusion axis reverses the sign by 250 labels (Section 5.2). The explanation is a quantity–quality trade-off: when almost nothing passes a hard threshold, softly including borderline pseudo-labels supplies a scarce gradient signal that is, on balance, better than none; once enough confident pseudo-labels exist, the extra borderline ones are net noise. The precision–coverage curves of Figure 5 show the same trade-off from the model side. Practitioners should therefore

treat soft inclusion as a knob to open at the smallest budgets and prefer a distribution-fitted shape when scarcity is extreme.

Curriculum and dataset difficulty. The per-class curriculum that drives the CIFAR-10 gains hurts on SVHN, where a fixed threshold is already reliable, and the threshold-trajectory diagnostic (Figure 11) shows why: the curriculum only has work to do when classes learn at visibly different rates. The lesson is that SSL design choices are not universally transferable; the appropriate recipe depends on how hard the unlabeled data is to classify confidently.

Other honest observations. Mean Teacher underperforms even supervised-only at 40 and 250 labels. Its unthresholded MSE consistency to an early, unreliable teacher reinforces errors without the confidence filtering the FixMatch family applies. The original Mean Teacher paper [12] uses a consistency-weight ramp-up schedule to attenuate the unsupervised loss while the teacher is still random; we deliberately ran the vanilla form with a fixed $\lambda_u = 1$ across all methods to keep the comparison uniform, and we treat the resulting underperformance as a known failure mode of un-ramped consistency under extreme scarcity rather than as a refutation of Mean Teacher in general. FixMatch’s fixed threshold becomes counter-productive at 4000 labels (below supervised-only): once enough labels are available, hard-thresholded high-confidence pseudo-labels admit a steady trickle of confident errors that supervised training avoids. Every method shows substantial seed variance at the smallest budgets (std up to ≈ 4 pp) because the identity of the few labeled examples is itself a major source of variability—a data-related challenge in its own right that any extreme-scarcity result must report.

Limitations. We collect the main caveats here so that readers can locate them without reading between the lines of earlier sections:

1. Compute budget. Every run uses 4000 optimization steps—orders of magnitude fewer than the original SSL papers—so absolute accuracies sit well below fully converged numbers and the magnitude of component effects could shift under longer training. The qualitative ordering we rely on (curriculum and soft weighting help under scarcity; agreement does not) is consistent across seeds and budgets, but we cannot guarantee it would persist at, say, 2^{20} steps. We deliberately frame the negative result on agreement within this compute-constrained regime.
2. Datasets and backbone. We evaluate on CIFAR-10, SVHN, and CIFAR-100 with a single Wide ResNet-28-2 backbone. STL-10 and harder image domains (medical, satellite, and industrial inspection) may give different component rankings; the SVHN and CIFAR-100 results (Section 5.6) already show that the curriculum component’s sign depends on dataset difficulty and on the density of its per-class statistics, including an outright failure mode at 100 classes under this budget.
3. Number of seeds. We use three seeds per cell, which is on the lower end for tight error bars. The variance is largely driven by which labeled examples each seed draws—itsself a data-related challenge—and we report all seeds explicitly so the noise is visible.
4. Baselines. We re-train supervised-only, Pseudo-Label, Mean Teacher, FixMatch, FlexMatch, and SoftMatch under the identical 4000-step budget. We do not include FreeMatch, whose self-adaptive thresholding is more sensitive to schedule length; reproducing it faithfully under a budget that does not match its published schedule would risk degrading that baseline unfairly, so we compare against the closest matched methods instead.
5. Hyperparameters. All hyperparameters ($\tau = 0.95$, $\gamma = 10$, $\kappa = 0.6$, $\rho = 0.5$, $\lambda_u = 1$, $m = 0.999$) are fixed across every dataset, budget, and seed; we did not perform per-budget tuning, which would risk overfitting to the regimes we evaluate. The soft-

weight sharpness γ is the parameter most likely to affect the quantity–quality trade-off documented in Section 5.2; a controlled sensitivity sweep is a natural follow-up.

6. EMA decay vs. schedule length. We use $m = 0.999$, the FixMatch default for approximately 1M-step training. Under 4000 steps, this is a slowly moving teacher, which we verified empirically improves over a faster decay; methods relying more heavily on the teacher (Mean Teacher) might benefit from a schedule-adapted m .

7. Conclusions

We presented CCM, a Confidence-Calibrated Consistency Matching method for semi-supervised image classification, together with a unified weighting view that locates it—and FixMatch, FlexMatch, and SoftMatch—in a single design space, and a controlled study of which SSL ingredients help under extreme label scarcity. This study’s clearest finding is about the weighting shape: at the hardest regime we examined (40 labels on CIFAR-10), the two smooth-weighting designs finished first and second—SoftMatch, whose shape was fitted to the confidence distribution, ahead of CCM’s simpler fixed-sharpness sigmoid ($p = 0.049$), with CCM in turn significantly ahead of the hard-threshold FlexMatch it extends ($p = 0.001$). Because CCM changes exactly one axis of FlexMatch’s design, this pair of significant gaps cleanly attributes the extreme-scarcity gains to smooth inclusion itself, with shape fitting adding a further increment; at 4000 labels, the ordering narrows and reverses numerically, and on the easier SVHN dataset, the curriculum-based methods as a group fall behind a fixed threshold. We also reported a negative result: in our setting, cross-view agreement—an intuitive confirmation-bias remedy—neither helps as an admission gate nor as a reliability reweighting, and reweighting consistently hurts; our diagnostics suggest a plausible reason (agreement is informative but not informative enough at this scarcity level), though we do not claim this conclusion transfers to other budgets or training regimes. Our overall takeaway is modest: under extreme label scarcity and a tight compute budget, softly admitting borderline pseudo-labels helps—by two independent designs’ testimony—intuitive stability-based filtering does not, and the right defaults can differ from those tuned on easier or larger-budget settings. All code, configuration files, per-run logs, and the manuscript-generation pipeline are released for reproduction on a single commodity GPU.

Author Contributions: Conceptualization, K.-S.S. and D.-H.W.; methodology, K.-S.S. and H.-G.P.; software, K.-S.S. and H.-G.P.; validation, K.-S.S. and D.-H.W.; formal analysis, K.-S.S. and D.-H.W.; investigation, K.-S.S. and D.-H.W.; resources, K.-S.S.; data curation, H.-G.P.; writing—original draft preparation, D.-H.W.; writing—review and editing, K.-S.S. and D.-H.W.; visualization, H.-G.P.; supervision, K.-S.S. and H.-G.P.; project administration, K.-S.S.; funding acquisition, D.-H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by Wonkwang University in 2025.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All code, configuration files, per-run logs, figures, and the manuscript-generation pipeline are released as an open-source repository, enabling full reproduction on a single commodity GPU. The CIFAR-10 and SVHN datasets are publicly available from their original sources.

Acknowledgments: The authors thank the maintainers of the FixMatch and FlexMatch reference implementations and the curators of the CIFAR-10 and SVHN datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SSL	Semi-Supervised Learning
CCM	Confidence-Calibrated Consistency Matching (proposed method)
EMA	Exponential Moving Average
ECE	Expected Calibration Error
RNG	Random Number Generator
SGD	Stochastic Gradient Descent
AMP	Automatic Mixed Precision
pp	percentage points

References

1. Su, J.; Luo, Z.; Lian, S.; Lin, D.; Li, S. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Med. Image Anal.* **2024**, *94*, 103111. [[CrossRef](#)] [[PubMed](#)]
2. Chen, H.; Ru, J.; Long, H.; He, J.; Chen, T.; Deng, W. Semi-supervised adaptive pseudo-label feature learning for hyperspectral image classification in internet of things. *IEEE Internet Things J.* **2024**, *11*, 30754–30768. [[CrossRef](#)]
3. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [[CrossRef](#)]
4. Yang, X.; Song, Z.; King, I.; Xu, Z. A survey on deep semi-supervised learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 8934–8954. [[CrossRef](#)]
5. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
6. Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; Shinozaki, T. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021.
7. Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
8. Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; Savvides, M. SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
9. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.E.; McGuinness, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020. [[CrossRef](#)]
10. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
11. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016.
12. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
13. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
14. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A holistic approach to semi-supervised learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
15. Zhou, S.; Tian, S.; Yu, L.; Wu, W.; Zhang, D.; Peng, Z.; Zhou, Z. Growth threshold for pseudo labeling and pseudo label dropout for semi-supervised medical image classification. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107777. [[CrossRef](#)]
16. Li, Z.; Ding, Y.; Wei, M.; Xu, X. Open-world barely-supervised learning via augmented pseudo labels. *Electron. Res. Arch.* **2024**, *32*, 5804–5818. [[CrossRef](#)]
17. Jing, K.; Ma, H.; Zhang, C.; Shang, Z.; Zha, Z.; Wang, X.; Pang, W. Confidence matters: Balanced pseudo-label selection for semi-supervised few-shot learning. *IEEE Trans. Multimed.* **2026**, 1–14. [[CrossRef](#)]

18. Kang, M.; Seo, S.; Min, M. PC-Match: Semi-supervised learning with progressive contrastive and consistency regularization. *IEEE Access* **2025**, *13*, 44543–44554. [[CrossRef](#)]
19. Wu, H.; Sun, J.; Chen, Q. Dual-training-based semi-supervised learning with few labels. *Appl. Sci.* **2024**, *14*, 4993. [[CrossRef](#)]
20. Zhang, T.; Luo, Y.; Wang, X.; Kwong, S. Semi-supervised segmentation on medical images with pseudo label calibration and uncertainty estimation. *Neural Netw.* **2025**, *189*, 107510. [[CrossRef](#)] [[PubMed](#)]
21. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017.
22. Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E.D.; Goodfellow, I.J. Realistic evaluation of deep semi-supervised learning algorithms. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018.
23. Wang, Y.; Chen, H.; Fan, Y.; Sun, W.; Tao, R.; Hou, W.; Wang, R.; Yang, L.; Zhou, Z.; Guo, L.Z.; et al. USB: A unified semi-supervised learning benchmark for classification. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks, New Orleans, LA, USA, 28 November–9 December 2022.
24. Zagoruyko, S.; Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*; BMVA Press: Malvern, UK, 2016. [[CrossRef](#)]
25. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. RandAugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020. [[CrossRef](#)]
26. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features From Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
27. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading digits in natural images with unsupervised feature learning. In Proceedings of the NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 16–17 December 2011.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.