A. Appendix

A.1. Additional Details for Table 1

Detailed dataset setting is in Table 1. We provide additional configuration details specifically for the Food-101 dataset in Table 2. Settings for the remaining datasets follow USB [10]. For the Food-101 dataset, we use a ViT-base model [5] with a patch size of 16 and an image size of 224.

Table 1. Details of Datasets

Dataset	#Classes	Fine-grained?	#Labeled
Food-101 [1]	101	✓	404 / 1010
CIFAR-100 [7]	100	✓	200 / 400
Semi-Aves [9]	200	✓	5959
STL-10 [3]	10	×	40 / 100
EuroSAT [6]	10	×	20 / 40

Table 2. Additional Hyper-parameters for Table 2

Dataset	Food-101
Image Size	224
Model	ViT-B-P16-224
Weight Decay	0.03
Layer Decay Rate	0.75
LR Scheduler	$\eta = \eta_0 \cos\left(\frac{7\pi k}{16K}\right)$
Weak Augmentation	Random Crop, Random Horizontal Flip
Strong Augmentation	RandAugment [4]

A.2. Pseudo-Code

The pseudo-code for our method is shown in Pseudo-code 1

A.3. Additional Study

We hypothesize that our method boosts SSL performance through two key reasons: (1). Aligning text embeddings of ground-truth labels with visual representations helps the model capture subtle visual differences through textual cues, improving stability and robustness during initial training. (2) Even incorrect pseudo-labels can be beneficial if their semantics are close to the ground truth, guiding visual representations toward more accurate clusters and enhancing generalization. Thus, we study the impact on initial model and benefit under wrong pseudo-labels.

Impact to Initial Model. We first investigate how aligning ground-truth label names with visual representations improves the initial model. To analyze this, we assume the initial model is trained solely on labeled images using supervised learning. We then compare the performance of standard supervised training with supervised training enhanced by class-aware contrastive learning, as shown in Table 3. The results demonstrate that incorporating class-aware contrastive loss enhances supervised training performance. This improvement occurs because aligning text embeddings with visual embeddings strengthens the model's ability to learn more effective visual representations.

Pseudo-code 1: SemiVisBooster

```
class SemiVisBooster:
  def __init__(self, label_names):
    # [Class#, text_embs_dims]
    self.text_embs_bank = LLM(label_names).detach()
    # Other initializations ...
  def train_one_batch(self, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}_w$, $\mathcal{U}_s$):
    # $\mathcal{X}$,$\mathcal{Y}$:labeled images and labels
    # $\mathcal{U}_w$,$\mathcal{U}_s$:weak and strong augmented
    #        unlabeled images
    outputs = self.model(torch.cat([$\mathcal{X}$, $\mathcal{U}_w$, $\mathcal{U}_s$])
    logits_$\mathcal{X}$, logits_$\mathcal{U}_w$, logits_$\mathcal{U}_s$ = outputs["logits"]
    feats_$\mathcal{X}$, feats_$\mathcal{U}_w$, feats_$\mathcal{U}_s$ = outputs["feats"]
    # =========Base SSL loss==========
    # mask: pseudolabel selection mask from base SSL
    $\mathcal{L}_s$, $\mathcal{L}_u$, mask, pseudolabels = baseSSL(...)
    # ==Class-aware Contrastive loss==
    text_embs_bank = self.proj(self.text_embs_bank)
    $\mathcal{L}_{\text{TEDS}}$ = TEDS(text_embs_bank)
    V_embs = torch.cat([feats_X, feats_$\mathcal{U}_s$[mask]])
    labels = torch.cat([$\mathcal{Y}$, pseudolabel[mask]])
    T_embs = text_embs_bank[labels]
    $\mathcal{L}_c$ = Class-aware_CL(V_embs, T_embs, labels)
    # $\lambda_u$, $\lambda_c$ and $\lambda_t$ are loss weights
    $\mathcal{L}_{total} = \mathcal{L}_s + \lambda_u * \mathcal{L}_u$ + $\underline{\lambda_c * \mathcal{L}_c + \lambda_t * \mathcal{L}_{\text{TEDS}}}$
    # Backpropagation ...
```

Table 3. **Impact on initial model**: aligning ground-truth label names with visual representations enhances the initial model.

Method	Food-101	
	404	1010
Supervised	27.3	47.3
Supervised + Class-aware Contrastive	30.9	50.2

Table 4. **Benefit under wrong pseudo-label**: Partial semantic alignment, even with incorrect pseudo-label names, enhances representation learning.

Method	Food-101	
	404	1010
Generated pseudo-label accuracy	27.3	47.3
FixMatch_F	35.8	60.3
FixMatch_F + Class-aware Contrastive	35.6	65.3

Benefit under wrong pseudo-label To fairly evaluate the benefit of class-aware contrastive learning under incorrect pseudo-labels, it is necessary to ensure consistent pseudo-label accuracy across comparisons. This consistency allows for a direct comparison between standard SSL methods and those incorporating class-aware contrastive learning. However, maintaining such consistency is challenging because pseudo-label accuracy in SSL methods dynamically changes during training. In traditional SSL methods [8], although gradients are not backpropagated to update the pseudo-label generator, the generator shares weights with the in-training model, causing pseudo-label predictions to evolve at each step. Advanced SSL methods, such as FlexMatch [12], FreeMatch [11], and SoftMatch [2], use dynamic thresholding for pseudo-label sampling, further complicating the effort to ensure consistent pseudo-label accuracy. To address this issue, we introduce FixMatch_F, an extension of FixMatch [8]. First, FixMatch_F adopts a fixed confidence threshold for pseudo-label selection to ensure consistency in the labeling criteria. Second, we pre-train the model on supervised data and freeze it during pseudo-label generation. This prevents the model from updating during SSL training, ensuring that pseudo-label accuracy remains consistent throughout the process.

As shown in Table 4, class-aware contrastive loss does not improve SSL performance when pseudo-label accuracy is very low, such as 27.3%. This limitation arises because the alignment of pseudo-label names with visual embeddings relies on semantic similarity between the two. When pseudo-labels are highly inaccurate, the semantic information they convey is incorrect, offering no benefit to the model. However, as pseudo-label accuracy improves, the performance impact of class-aware contrastive learning becomes significant. This improvement occurs because even partial semantic alignment between pseudo-label names and visual embeddings enhances the model's ability to learn accurate visual representations, leading to better overall performance.

Robustness of TEDS. The accuracy gain from TEDS decreases as more labeled data becomes available, because the model can learn subtle visual differences from labeled data (Table 5). The visual-text alignment loss encourages mutual enhancement: better visual representations lead to more distinct text embeddings, and vice versa. With more labels, the challenges naturally diminish. Our target is to address fine-grained challenges under limited labels, where the visual features are not distinct. So, TEDS is important. In addition, with more labeled data, TEDS still provides benefits and does not degrade performance. This confirms that TEDS is most beneficial under low-label regimes.

Table 5. TEDS performance gain along with labeled images

# Labeled Images	Accuracy gain	Average of cosine similarity W/O TEDS	W/ TEDS
404	4.96	0.081	0.067
1010	0.39	0.071	0.068
2020	0.43	0.067	0.052

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461. Springer, 2014. 1

[2] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[3] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, Fort Lauderdale, FL, USA, 2011. PMLR. 1

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1

[7] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[8] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[9] Jong-Chyi Su and Subhransu Maji. The semi-supervised inaturalist-aves challenge at fgvc7 workshop. *arXiv preprint arXiv:2103.06937*, 2021. 1

[10] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022. 1

[11] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive

thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 2

[12] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021. 2