PAF: Perturbation-Aware Filtering for Open-Set Semi-Supervised Learning

Supplementary Material

A. Empirical Lipschitz Constant Estimation

In the main text, our theoretical analysis assumes that every logit $f_{\theta,k}$ is β-Lipschitz [5] in the input. To verify that this assumption is empirically reasonable, we measure the effective Lipschitz constant of our trained WideResNet-28×2 models. For each independently trained model (three random seeds), we freeze the weights and apply a single-step power iteration to estimate the largest singular value $\|W^{(l)}\|_2$ of every convolutional or fully-connected layer. The global constant is then approximated by the product of these layer-wise spectral norms:

$$\beta_{\text{emp}} \;=\; \prod_{l=1}^{L} \|W^{(l)}\|_2. \tag{1}$$

Discussion. Expressing β_{emp} in its base-10 logarithmic form (i.e., reporting $\log_{10}(\beta_{\text{emp}})$, which falls between 3.30 and 3.64) makes the values compact and easy to compare. In this scale, a difference of 1 means that the underlying Lipschitz constant changes by one order of magnitude (a factor of 10).

As shown in Table I, β_{emp} consistently falls within the 10^3–10^4 range across all seeds. This is approximately two orders of magnitude smaller than the $10^5 - 10^6$ values typically observed in vanilla ResNet models without spectral control. This confirms that our networks remain within a reasonable Lipschitz regime even without explicit regularization, and validates the β-Lipschitz assumption used in our theorem analysis.

Table I. Empirical β on three random seeds for WideResNet-28×2.

Seed	β_{emp}	Max layer-norm	Min layer-norm
0	3.51 (3.2×10^3)	2.74	0.86
1	3.30 (2.0×10^3)	2.79	0.75
2	3.64 (4.4×10^3)	3.06	0.81

B. Supplementary Theoretical Proofs

In this section, we provide the theoretical proofs supporting the statements used in the main text. Our analysis focuses on local neighborhoods induced by semantic-preserving perturbations, where the classifier behaves smoothly and decision scores do not exhibit abrupt changes. All results below are conditional on these mild regularity properties.

Setup. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$ be the logit map with a linear classifier applied to the penultimate representation $\varphi(\mathbf{x}) \in \mathbb{R}^d$:

$$f_\theta(\mathbf{x}) = W\,\varphi(\mathbf{x}) + b, \qquad W \in \mathbb{R}^{K\times d}.$$

Let $\sigma(\cdot)$ denote the softmax and define the top-1 confidence

$$s(\mathbf{x}) \;=\; \max_k \sigma_k\big(f_\theta(\mathbf{x})\big),$$

and the decision margin:

$$m(\mathbf{x}) = f_{\theta,y(\mathbf{x})}(\mathbf{x}) - \max_{k\neq y(\mathbf{x})} f_{\theta,k}(\mathbf{x}),$$
$$y(\mathbf{x}) = \arg\max_k f_{\theta,k}(\mathbf{x}).$$

We focus on inputs with a unique predicted class so that $s(\mathbf{x}) = \sigma_{y(\mathbf{x})}(f_\theta(\mathbf{x}))$ is differentiable in a neighborhood of $\mathbf{x}$.

Assumption. Semantic-preserving perturbations and augmented views of $\mathbf{x}$ typically remain in a locally stable region of the classifier, where both the predicted class and the margin vary smoothly along short interpolation paths. Motivated by this empirical observation, we impose the following mild condition.

For $u \in \{\mathbf{x}, \varphi\}$ and any pair of classes $i \neq k$, the logit difference

$$g_{i,k}(u) = f_{\theta,i}(u) - f_{\theta,k}(u)$$

is β-Lipschitz [5] in u, i.e.,

$$\|\nabla_u g_{i,k}(u)\|_2 \;\le\; \beta.$$

Under the linear classifier $f_\theta(\mathbf{x}) = W\varphi(\mathbf{x}) + b$, a valid representation-space choice is:

$$\beta \;:=\; \max_{i\neq k} \|w_i - w_k\|_2.$$

Absolute numerical constants are absorbed into β.

Lemma 1 (Gradient–margin bound). For any $\mathbf{x}$, the gradient of the top-1 confidence score satisfies

$$\big\|\nabla_{\mathbf{x}}\, s(\mathbf{x})\big\|_2 \;\;\le\;\; \beta\,(K-1)\,e^{-m(\mathbf{x})}.$$

Proof. Let $p = \sigma(f_\theta(\mathbf{x}))$ and $s = p_y$ with $y = y(\mathbf{x})$. The softmax Jacobian yields

$$\nabla s = \sum_{i=1}^{K} \frac{\partial s}{\partial f_{\theta,i}}\,\nabla f_{\theta,i} = s \sum_{k\neq y} p_k\,\nabla\big(f_{\theta,y} - f_{\theta,k}\big).$$

Using the Lipschitz assumption, $\|\nabla(f_{\theta,y} - f_{\theta,k})\|_2 \le \beta$ holds for all $k \neq y$, giving:

$$\|\nabla s\|_2 \;\le\; s\sum_{k\neq y} p_k\,\beta \;=\; \beta\, s(1-s) \;\le\; \beta\,(1-s).$$

Write $a = f_{\theta,y}(\mathbf{x})$ and $b_k = f_{\theta,k}(\mathbf{x})$ for $k \neq y$. Then

$$1-s = \frac{\sum_{k\neq y} e^{b_k-a}}{1+\sum_{k\neq y} e^{b_k-a}} \;\le\; \sum_{k\neq y} e^{b_k-a} \;\le\; (K-1)e^{-m(\mathbf{x})},$$

because $b_k - a \le -m(\mathbf{x})$ for all $k \neq y$. Combining terms yields

$$\big\|\nabla_{\mathbf{x}}\, s(\mathbf{x})\big\|_2 \;\le\; \beta(K-1)\, e^{-m(\mathbf{x})}.$$

□

Theorem 1 (Variance upper bound). Let δ be a random perturbation supported on the local semantic-preserving region

$$\left\{\delta:\; \begin{array}{l} \|\delta\|_2 \le \xi, \\ y(\mathbf{x}+t\delta) = y(\mathbf{x})\;\forall\, t\in[0,1], \\ m(\mathbf{x}+t\delta) \ge m(\mathbf{x}) - \varepsilon,\;\forall\, t\in[0,1] \end{array}\right\},$$

where $\varepsilon \ge 0$ allows for mild non-monotonicity of the margin along short interpolation segments. The constant factor induced by ε is absorbed into β in the final bound. Write

$$\sigma^2 \;:=\; \mathbb{E}\big[\|\delta\|_2^2\big].$$

Then, for any $\mathbf{x}$,

$$\mathrm{Var}_\delta\big[s(\mathbf{x}+\delta)\big] \;\le\; \beta^2(K-1)^2\, e^{-2m(\mathbf{x})}\,\sigma^2.$$

Proof. Using the fundamental theorem of calculus along the segment $\{\mathbf{x}+t\delta : t\in[0,1]\}$,

$$s(\mathbf{x}+\delta) - s(\mathbf{x}) = \int_0^1 \nabla s(\mathbf{x}+t\delta)^\top \delta\, dt.$$

Let $c = s(\mathbf{x})$. Then

$$\mathrm{Var}_\delta[s(\mathbf{x}+\delta)] \;\le\; \mathbb{E}_\delta\big[(s(\mathbf{x}+\delta)-c)^2\big].$$

Applying Cauchy–Schwarz and the local regularity conditions,

$$\mathrm{Var}_\delta[s(\mathbf{x}+\delta)] \le \Big(\sup_{t\in[0,1]} \|\nabla s(\mathbf{x}+t\delta)\|_2\Big)^2\, \mathbb{E}_\delta\big[\|\delta\|_2^2\big].$$

Using the previously established gradient–margin bound along the stable segment gives

$$\|\nabla s(\mathbf{x}+t\delta)\|_2 \le \beta(K-1)e^{-m(\mathbf{x}+t\delta)} \le \beta(K-1)e^{-m(\mathbf{x})},$$

for all t, up to constants absorbed into β. Therefore,

$$\mathrm{Var}_\delta[s(\mathbf{x}+\delta)] \;\le\; \big(\beta(K-1)e^{-m(\mathbf{x})}\big)^2\,\sigma^2.$$

□

Proposition 1 (Representation Confidence Coupling). Let $\mathbf{x}'$ be a semantic-preserving view of $\mathbf{x}$ with $y(\mathbf{x}') = y(\mathbf{x})$. We assume the interpolation path in representation space,

$$\gamma(t) = (1-t)\varphi(\mathbf{x}) + t\,\varphi(\mathbf{x}'), \quad t\in[0,1],$$

lies in a locally stable region where the predicted class remains unchanged and the margin does not decrease beyond a small tolerance. Then

$$\big|s(\mathbf{x}) - s(\mathbf{x}')\big| \;\le\; \beta\,(K-1)\,e^{-\,m(\mathbf{x})}\,\big\|\varphi(\mathbf{x}) - \varphi(\mathbf{x}')\big\|_2.$$

Proof. Let $h(\varphi) := s(\varphi)$ and write $\varphi_0 = \varphi(\mathbf{x})$, $\varphi_1 = \varphi(\mathbf{x}')$. Along the straight-line interpolation $\gamma(t) = (1-t)\varphi_0 + t\varphi_1$,

$$s(\mathbf{x}') - s(\mathbf{x}) = h(\varphi_1) - h(\varphi_0) = \int_0^1 \nabla_\varphi h(\gamma(t))^\top \gamma'(t)\, dt.$$

Hence,

$$\big|s(\mathbf{x}')-s(\mathbf{x})\big| \le \left(\sup_{t\in[0,1]} \big\|\nabla_\varphi s(\gamma(t))\big\|_2\right)\, \big\|\varphi(\mathbf{x}')-\varphi(\mathbf{x})\big\|_2.$$

For the linear classifier $f_\theta = W\varphi + b$, we have $\nabla_\varphi f_{\theta,k} = w_k$, and by the Lipschitz assumption, $\|\nabla_\varphi(f_{\theta,y} - f_{\theta,k})\|_2 \le \beta$ for all $k \neq y$. Repeating the proof of Lemma 1 in representation space yields

$$\|\nabla_\varphi s(\gamma(t))\|_2 \le \beta(K-1)e^{-m(\gamma(t))}.$$

By local stability, $m(\gamma(t)) \ge m(\mathbf{x}) - \varepsilon$ for all $t\in[0,1]$, hence

$$\|\nabla_\varphi s(\gamma(t))\|_2 \le e^{\varepsilon}\,\beta(K-1)e^{-m(\mathbf{x})}.$$

Absorbing the multiplicative constant e^{ε} into β and combining the bounds, we obtain

$$\big|s(\mathbf{x}) - s(\mathbf{x}')\big| \;\le\; \beta\,(K-1)\,e^{-m(\mathbf{x})}\,\big\|\varphi(\mathbf{x}) - \varphi(\mathbf{x}')\big\|_2,$$

which completes the proof. □

C. Implementation Details

Detailed Dataset Settings. We organize the dataset details into two complementary settings: internal OOD and external OOD scenarios. For internal OOD scenario, we generate ID and OOD data by partitioning the same dataset into seen (ID) and unseen (OOD) classes, so both types of samples originate from a common source yet differ semantically.

- **MNIST.** MNIST [4] contains 10 classes ranging from digit "0" to digit "9", comprising 60,000 training images of size 28×28. We select 10 images from each of the classes "0" to "5" as the labeled set, and 30,000 images from all classes as the unlabeled data.

- **CIFAR-10.** CIFAR-10 [3] includes 10 classes, each containing 6,000 images of size 32×32. We designate the animal classes as seen classes and the remaining classes as unseen. A total of 2,400 labeled images (400 from each seen class) and 20,000 unlabeled images (randomly sampled from all classes) are selected for training.
- **CIFAR-100.** CIFAR-100 is an extension of CIFAR-10 with 100 classes [3]. The first 50 classes are used as seen classes and the remaining 50 as unseen classes. The labeled set contains 5,000 images, obtained by sampling 100 images per seen class, while the unlabeled set consists of 20,000 images randomly selected across all classes.
- **TinyImageNet.** TinyImageNet is a subset of ImageNet [1], consisting of 100,000 images from 200 classes. All images are resized to 32×32. The first 100 classes are considered as seen, and the remaining as unseen. We select 100 images from each seen class to build the labeled set, and 40,000 images are randomly sampled from all classes to build the unlabeled set.

Furthermore, for external OOD scenario, the OOD samples come from different datasets, real-world or synthetic, and are explicitly injected into the unlabeled data to simulate distributional shifts.

- **LSUN.** The Large-scale Scene Understanding (LSUN) [7] dataset contains tens of millions of high-resolution images covering diverse indoor and outdoor scenes. In our experiments, LSUN serves as an external OOD dataset to create a realistic open-set scenario. Specifically, 10,000 images are sampled from LSUN to serve as OOD data within the unlabeled dataset.
- **Synthetic Noise.** The Gaussian Noise and Uniform Noise datasets are synthetic perturbation-based OOD benchmarks provided by the official implementation of T2T [2]. In our experiments, they are used as external OOD datasets to construct the open-set scenario within the unlabeled data.
- **TinyImageNet.** Following the setting in T2T [2], we sample 10,000 images from TinyImageNet [1] as external OOD data to be included in the unlabeled set.

We additionally evaluate our approach on the CUB-200-2011 dataset [6], a widely used fine-grained bird classification benchmark comprising 11,788 images from 200 visually similar bird species. The dataset is characterized by subtle inter-class differences, making it a challenging benchmark for fine-grained classification. Following the standard protocol, we use 5,994 images for training and 5,794 for testing.

Hardware. All experiments were conducted on a workstation equipped with a single NVIDIA 3090-24GB GPU and dual Intel Xeon Gold 5220R CPUs.

Software. All experiments are conducted on Ubuntu 22.04 LTS with Python 3.10.13, using PyTorch 2.1.0 compiled with CUDA 12.2, together with torchvision 0.16.0 and cuDNN 8.9.

Training Hyper-parameters. We summarize all hyper-parameters used in our experiments in Table II.

Table II. Training hyper-parameters used throughout the experiments.

Parameter	Value
Backbone	WideResNet-28-2
Optimizer	SGD
Momentum	0.9
Learning rate	0.03 (cosine annealing)
Weight decay	5×10^{-4}
Labeled data Batch size	64
Unlabeled data Batch size	128
Stage 1 training steps	50,000
Stage 2 training steps	200,000
Eval steps	1,000
Filtering frequency	Every 20 epochs (Every 20,000 steps)

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 3

[2] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting OOD data with cross-modal matching for open-set semi-supervised learning. In *ICCV*, pages 8290–8299. IEEE, 2021. 3

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009. 3

[4] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*, 2, 2010. 2

[5] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *JMLR*, 5:669–695, 2004. 1

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 3

[7] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 3