Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation
Supplementary Material

Lihe Yang[1] Lei Qi[2] Litong Feng[3] Wayne Zhang[3] Yinghuan Shi[1]
[1]Nanjing University [2]Southeast University [3]SenseTime Research

https://github.com/LiheYoung/UniMatch

A. How about Removing Image-Level Strong Perturbations?

Perturbation Level	92	183	366	732	1464	1/16	1/8	1/4
Image Level Alone (FixMatch)	63.9	73.0	75.5	77.8	79.2	74.1	75.9	76.4
Feature Level Alone	66.0	69.6	74.0	77.3	78.9	73.7	74.5	76.4
Unified Levels (UniPerb)	72.0	75.8	77.5	79.3	80.1	76.0	76.9	76.6

Table 1. Results (%) of only using *single* perturbation level, either image-level perturbations (original FixMatch) or feature-level perturbations. These results are obtained from the Pascal dataset with DeepLabv3+ and ResNet-101. We also provide results of our UniPerb (Unified Levels) as a reference, which unifies the two different levels of perturbations.

It has almost become a primary concern and a common practice in various semi-supervised settings to seek proper image-level strong perturbations first. Nevertheless, this process requires many time-consuming trials and delicate selection of different combinations. To make matters worse, in some domain-specific tasks, such as medical image analysis, it is challenging for most practitioners to figure out appropriate ones. Therefore, a natural question raises: could we replace image-level strong perturbations in FixMatch with a simple channel dropout perturbation at the feature level?

To validate this, we make a modification to original FixMatch that, input images are not processed by any strong data augmentation, but their features are perturbed by a channel dropout. It can be observed from Table 1 that in most cases, feature-level perturbation alone can indeed perform on par with original image-level strong perturbations in FixMatch, merely slightly inferior. Hence, we believe that such simple but universal feature perturbations may serve as a promising supplement, when image-level strong perturbations fail to work in some rarely explored scenarios.

B. Dual-Stream Feature-Level Perturbations

Method	92	183	366	732	1464	1/16	1/8	1/4
Single-Stream FP (UniPerb)	72.0	75.8	77.5	79.3	80.1	76.0	76.9	76.6
Dual-Stream FP	**73.4**	**77.1**	**78.5**	**79.6**	**80.4**	**76.2**	**77.0**	**76.8**

Table 2. Effectiveness of dual-stream perturbations at the feature level (%). FP here denotes feature perturbation. Same as the single-stream FP (UniPerb), the dual-stream FP also contains one stream for image-level strong perturbations.

The technique of dual-stream perturbations has been proved to be highly beneficial at the image level. Certainly, we wish to check its effectiveness at the feature level. Thus, we attempt to strengthen our proposed UniPerb via performing twice parallel channel dropout on the extracted features. The dual perturbed features are then sent into the decoder to produce two final predictions for learning. The results of dual-stream feature-level perturbations are reported in Table 2. Our UniPerb

can be further boosted via maintaining dual feature perturbation streams. As discussed in Section 3.3 of our main paper, we conjecture that dual random perturbations on the same features can also be considered to produce a pair of positive views, thereby harvesting the merits of contrastive learning. Despite the effectiveness, we decide not to conduct dual-stream feature perturbations in our main approach, because the current version is powerful enough, and we hope to avoid additional computational burden during training.

C. More Image-Level Perturbation Streams

Number of labeled images	Number of image-level perturbation streams						
	1 (FixMatch)	2 (DusPerb)	3	4	5	6	7
High-quality set: 732	77.8	78.1	78.8	78.9	**79.1**	78.7	78.4
Blended set: 662 (1/16)	74.1	75.3	75.4	76.1	76.0	**76.7**	76.3

Table 3. The performance (%) change with respect to the number of image-level strong perturbation streams.

Here, the auxiliary feature-level stream is excluded, which means the perturbation space is completely constrained at the image level. Then, we progressively increase the number of image-level strong perturbation streams on the Pascal, and report the corresponding performance in Table 3.

The performance is steadily improved as the number of strong views is increased to a certain number. But if we continue to increase beyond it, then the performance might drop a little. The results indicate that, two or three strong views are already enough to fully probe the original image-level perturbation space. Excessive strong views might cause the model to struggle in learning every single view.

D. Limitations, Discussions, and Future Works

In our UniMatch, a confidence threshold is primarily set to suppress potentially incorrect pseudo labels. In some challenging scenarios, *e.g*. COCO, however, we observe that around 15% pixels are discarded during the learning course. Therefore, how to make full use of these uncertain pixels and meantime avoid error accumulation will be a promising direction to further facilitate current semi-supervised algorithms. This may also enable our model to be more robust to different thresholds.

Moreover, our framework, along with its precedents, such as the FixMatch serials [1–3,5] and UDA [4], heavily relies on the pseudo labeling quality on unlabeled images. In case yielded pseudo labels are poor, it would be hard for our semi-supervised learner to mine meaningful knowledge from unlabeled images. Therefore, if the class distribution is highly imbalanced, the model will be gradually biased to majority classes during training and pseudo labeling, making the minority classes worse and worse. In addition, on common benchmarks, the domain gap between labeled and unlabeled images is rarely considered. However, in real worlds, the abundant unlabeled images can not share exactly the same domain as labeled ones. The semi-supervised learner could benefit from more unlabeled images if domain shift is well addressed.

Last but not least, existing academic settings in semi-supervised classification/segmentation/detection prefer to restricting labeled images to an extremely low proportion, *e.g*. only providing 40 labels on the CIFAR-10 and 92 labeled images on the Pascal. Nevertheless, considering most real-world demands, it might be more practical to assume labeled images is in the tens of thousands, while unlabeled images are even more, might in millions. Actually, a prior work [6] already explored such a setting, but it is expected to be further improved, both in accuracy and training efficiency.

We leave the aforementioned four problems, namely 1) how to fully exploit uncertain pixels, 2) class imbalance in pseudo labeling, 3) domain shift in pseudo labeling, and 4) how to effectively benefit from millions of unlabeled samples together with considerable labeled ones, to our future works.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *ICLR*, 2020. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2

[3] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 2

[4] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 2

[5] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021. 2

[6] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 2