GENERALIZED PSEUDO-LABELING IN CONSISTENCY REGULARIZATION FOR SEMI-SUPERVISED LEARNING

Nikolaos Karaliolios, Florian Chabot, Camille Dupont, Hervé Le Borgne,
Quoc-Cuong Pham, Romaric Audigier

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

ABSTRACT

Semi-Supervised Learning (SSL) reduces annotation cost by exploiting large amounts of unlabeled data. A popular idea in SSL image classification is Pseudo-Labeling (PL), where the predictions of a network are used in order to assign a label to an unlabeled image. However, this practice exposes learning to confirmation bias. In this paper we propose Generalized Pseudo-Labeling (GPL), a simple and generic way to exploit negative pseudo-labels in consistency regularization, entailing minimal additional computational overhead and hyperpameter fine-tuning. GPL makes learning more robust by using the information that an image *does not* belong to a certain class, which is more abundant and reliable. We showcase GPL in the context of FixMatch. In the benchmark using only 40 labels of the CIFAR-10 dataset, adding GPL on top of FixMatch improves the error rate from 7.93% to 6.58%, and on CIFAR-100 with 2500 labels, from 28.02% to 26.85%.

***Index Terms*—** Semi-Supervised Learning, Pseudo-Labeling, Consistency Regularization

1. INTRODUCTION

Since Machine Learning algorithms have achieved high performances in benchmark image classification tasks with abundant labeled data [1, 2, 3, 4], focus has started to shift toward the more difficult case of scarce labeled data [5, 6, 7, 8, 9, 10, 11]. Commonly related to annotation cost, scarcity of labeled data is often paired up with a large amount of unlabeled data being available readily or at a low cost. This shift in focus led to the rise of the domain known as Semi-Supervised Learning (SSL) where the applied techniques combine Supervised Learning techniques using the labeled data, along with Unsupervised Learning ones, using the unlabeled data.

An important difficulty lies in efficiently exploiting the unlabeled data, while avoiding enforcing confirmation bias (CB) [12]. CB appears when an image is attributed to a certain class at a certain point of training, and the minimization of the loss function used in training enforces this belief of the network. In case of initial misclassification, confirmation bias is obviously detrimental to performance.

Pseudo-Labeling (PL), introduced in [7], consists in using the prediction of a neural network in order to assign a class to an unlabeled image, provided that the class score is above a certain threshold, fixed as a hyperparameter. This class, called the pseudo-label (pl) of the image, is then used in a supervised training strategy. Adopting a low threshold for attributing a pl leads to an increased degree of exploitation of unlabeled data, but introduces the risk of overexposing training to mislabeling, especially in the early stages of training when the predictions of the network tend to be less reliable. The trade-off between guarding against CB and increasing the degree of data exploitation is one of the main difficulties in this technique. PL improves on the supervised baseline using only labeled data, but needs a regularization loss function in order to do so, and, predictably, suffers from CB.

The Consistency Regularization (CR) exploits random augmentations, based on the intuition that while different (reasonable) augmentations of the same image belong to the same class, they do not necessarily belong to the same statistical distribution. This approach is used in [9, 10, 11], achieving a significantly better balance between data exploitation and CB. Combining it with PL further resulted in a large variety of new performing approaches [8, 10, 13].

In academic datasets and existing SSL benchmarks, classes are equally represented. However, in real-life applications this is in general not the case. Such unbalancing enforce CB against classifying images in the less represented classes. Therefore, assumptions on the distribution of class labels, as in [14] should be introduced with extreme prudence as they can significantly hinder generalizability. Our work improves any SSL approach that uses PL, without assumptions on class distribution, allowing a better adaptation to real-life settings.

We propose the Generalized Pseudo-Labeling (GPL) that uses both positive and negative pseudo labels (pls). We present a more general Teacher-Student (T-S) framework, where the Teacher provides the pls and the Student learns them. Negative pls consist in the information that a certain image *does not* belong to a certain class, materialized by low class-score predictions by the Teacher. The student in GPL

This work was funded by the project AVFS and made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council.

can thus learn that an image does not belong to a class, before knowing to which class it actually belongs. The challenges in applying GPL are two-fold: GPL introduces new hyperparameters (threshold for accepting negative pls and the weight of the corresponding loss); moreover, the loss function had to be carefully chosen (regression instead of negative log-loss) in order to avoid instabilities. The weight of the loss on negative pls is gauged on the loss on positive ones, resulting in minimal fine-tuning of the additional hyperparameter.

In order to investigate the sensitivity of performance with respect to class imbalance, we introduce a new benchmark on CIFAR-10 with 40 labels, where the unlabeled set corresponding to one class is reduced to 60% and show that our method is more robust with respect to class distribution.

A simple heuristic shows that using positive as well as negative pls feeds the network with more abundant and more reliable information, earlier in training. Our experiments confirm that the network is fed with a greater amount of information, which is also of better quality. As a consequence, learning converges faster and to a better score, and the learning curve is smoother. Our heuristics are confirmed since the purity of negative pls reaches high levels quite fast, and data coverage is significantly higher throughout training.

2. RELATED WORK

For a more complete presentation of the state of the art in the field of Semi-Supervised Learning, we refer the reader to [5]. FixMatch, introduced in [8], applied PL in CR, and has drawn a lot of attention in recent literature. The authors applied a variant of self-supervised techniques already present in the literature [9, 6, 11, 7] using the predictions on a weakly augmented version of the image as pls, in order to learn the class of the strong augmentation. FlexMatch [14], improved on FixMatch in the scarce label regime using an adaptive threshold, taking into account the number of positive pls attributed to each class and the maximal score for each class. However, this improvement depends crucially on the uniformity of the class distribution of the benchmark datasets (or at the very least on a good estimation thereof). Knowledge of the class distribution is unrealistic in a real-world SSL scenario with few labeled examples, and extrapolation from only a few labels would require an adaptation of the predicted class distribution during training which is not included in FlexMatch as it is. Our approach is agnostic of the class distribution, which makes it applicable to real-world SSL scenarios. Other recent improvements include Meta-Pseudo Labels [10], Selfmatch [15], Comatch [16], DASH [13], some of which actually contain a version of FixMatch as a submodule. Implementing GPL on top of these techniques is possible. As our method has the same level of generality as the original Pseudo-Labeling method of [7], in order to showcase it we need to choose a specific context for our experiments, which for this article is FixMatch [8], the standard baseline used by all successive improvements since its appearance.

Negative pseudo labels were used by [17] to learn with noisy label. The UPS method of [18] uses the setup of [7] with negative learning as in [17], but their application entails additional hyperparameters and was not applied on CR, resulting in low performance in the scarce label regime.

3. GENERALIZED PSEUDO-LABELING

The main observation behind negative pls is that, especially early in training, the prediction of a model that a given image *does not* belong to a class is bound to be more trustworthy than the prediction that the image *does* belong to a class. Experiments confirm this observation, as negative pl purity is high throughout training, and, moreover, all points have at least one negative pl practically at the beginning of training. For a classification problem with n classes, $n-1$ negative pls are equivalent to a positive one, since each image belongs to exactly one class.

By pure chance, a positive pl is accurate with probability only $\frac{1}{n}$. Attributing k negative pls, $1 \leq k < n$, is accurate with probability $\frac{n-k}{n}$, by pure chance. Reasonable thresholding can assure that k will not be comparable to n before the predictions of the model become accurate enough, assuring a better purity of negative pls and improved data exploitation. Providing the network with labeled examples and using consistency regularization naturally improves the accuracy of both positive and negative pls. However, negative pls are bound to be more abundant and more accurate to begin with, and this should not be expected to change in the presence of labeled examples. This heuristic assumes uniform class distribution, but this assumption can be lifted.

GPL thus consists in using the negative predictions of the Teacher, as well as the positive ones. A negative prediction is materialized by a low score prediction for the corresponding class. GPL is applied via the introduction of a new loss function allowing for negative pls to be exploited, leading to a better exploitation of images, without exposing the model to increased confirmation bias. As a result, convergence is faster, since more and better quality information is provided to the network earlier in training. GPL has the same scope as PL, that is it be applicable in every context where PL has been or can be applied.

We consider an SSL classification problem with $n \geq 3$ classes. The total train dataset $\mathcal{X}$ is split into $\{l_i\} = \mathcal{L}$, the set of labeled images, and $\{u_i\} = \mathcal{U}$, the set of unlabeled ones. All n classes are present in $y = \{y_i\} \in \mathcal{Y} = [\![0, n-1]\!]^{\#\mathcal{L}}$, the labels of images in $\mathcal{L}$.

Pseudo-Labeling uses a confidence threshold $\bar{\tau} \in (0.5, 1]$ to truncate scores and obtain the pseudo-labels in $\{0,1\}^n$ via the function $PL(s;\bar{\tau}) = PL(s) = \mathbb{1}(s \geq \bar{\tau})$.

3.1. Generalized Pseudo-Labeling

GPL attributes positive as well as negative pls thus obtaining the generalized pl of an unlabeled image. We introduce

the threshold $\underline{\tau} \in [0, \frac{1}{n})$ for accepting a negative pl. GPL is formalized by replacing the pseudo-labeling function GPL taking values in $\{-1, 0, 1\}^n$ given by

$$GPL(s; \bar{\tau}, \underline{\tau}) = GPL(s) = \mathbb{1}(s \geq \bar{\tau}) - \mathbb{1}(s \leq \underline{\tau}). \quad (1)$$

where 1 stands for a positive PL, -1 for a negative one, and 0 for absence of pl. For a given score vector s, the vector $GPL(s)$ can contain 0 or 1 positive pls, plus a certain number of negative ones, between 1 and $n-1$. This enables the model to learn earlier, even in the absence of a positive pl.

Since $\underline{\tau} < \frac{1}{n}$, thus no more than $n-1$ negative pls are attributed. If an image is attributed $n-1$ negative pls with scores close to $\underline{\tau}$, the score of the dominant class is close to $1-(n-1)\underline{\tau}$. Equating that to $\bar{\tau}$ leads to $\underline{\tau} = \frac{1-\bar{\tau}}{n-1}$, which determines $\underline{\tau}$ as a function of $\bar{\tau}$ and n, and thus remove one hyperparameter to determine.

3.2. Teacher-Student setup

In its full generality the Teacher-Student setup uses a pseudo-labeling strategy $\mathcal{P}$ and two networks with softmax output: the Teacher $\tilde{g} = \tilde{g}_{\theta_1}$ feeding its scores to the pseudo-labeling strategy, and the Student $f = f_{\theta_2}$, that learns the pls. Ignoring batching, the resulting Teacher-Student loss function is then:

$$\mathcal{L}_{T\text{-}S} = \mathbb{E}_{\mathcal{L}}\left[CE(\tilde{g}(\cdot), y)\right] + \lambda_u \mathbb{E}_{\mathcal{U}}\left[\Phi(f(\cdot), g(\cdot))\right], \quad (2)$$

where y are the ground truth labels, Φ is an adapted supervised loss function, $g = \mathcal{P} \circ \tilde{g}$, and λ_u is a weight. [7, 8, 14] are special cases of this setup where the Teacher and the Student are trained at the same time.

3.3. GPL loss function

In the general case of GPL, we implement the T-S setup of Eq. (2) with $\mathcal{P} = GPL$ and the corresponding unsupervised loss function $\Phi = L_{GPL}(\cdot)$ defined by

$$L_{GPL}(x) = \bar{\lambda} \sum_{\hat{y}_i = 1} \bar{\ell}(s_i) + \underline{\lambda} \sum_{\hat{y}_i = -1} \underline{\ell}(s_i). \quad (3)$$

Here, $\hat{y} = GPL(\tilde{g}_{\theta_1}(x))$ are the Generalized Pseudo-Labels and $s = f_{\theta_2}(x)$ is the output of the student for $x \in \mathcal{U}$. We introduce two functions $[0, 1] \to \mathbb{R}$, the strictly decreasing $\bar{\ell}$, and the strictly increasing $\underline{\ell}$. The factors $\bar{\lambda}$ and $\underline{\lambda}$ balance the loss of positive and negative pls. The function L_{GPL} reduces to CE when $\bar{\ell} = -\log$ and $\underline{\lambda} = 0$. Hence the loss is

$$\mathcal{L}_{GPL} = \mathbb{E}_{\mathcal{L}}\left[\mathrm{CE}(\tilde{g}(\cdot), y)\right] + \mathbb{E}_{\mathcal{U}}\left[L_{GPL}(f(\cdot), g(\cdot))\right], \quad (4)$$

where the factor λ_u of Eq. (2) is absorbed by $\bar{\lambda}$ and $\underline{\lambda}$.

Linearization of L_{GPL} in the neighborhood of a fully pseudo-labelized image gives, with s_p the score of the positive class and σ the vector of the scores of negative classes,

$$\mathrm{d}L_{GPL} = -\bar{\lambda} \frac{\bar{\ell}(s_p)}{\underline{\ell}^{-1}(\sigma)} \mathrm{d}s_p + \underline{\lambda} \frac{\mathrm{d}}{\mathrm{d}\sigma} \underline{\ell}(\sigma) \mathrm{d}\sigma. \quad (5)$$

	CIFAR-10-40	CIFAR100-2500
Fully Sup.*	4.62 ± 0.05	19.30 ± 0.09
ReMixMatch	19.10 ± 9.64	27.43 ± 0.31
DASH (RA)	13.22 ± 3.75	27.18 ± 0.21
DASH (CTA)	9.16 ± 4.31	27.85 ± 0.19
SelfMatch	6.81 ± 1.08	-
CoMatch	6.91 ± 1.39	-
FixMatch (RA)	13.81 ± 3.37	28.29 ± 0.11
FixMatch (CTA)	11.39 ± 3.35	28.64 ± 0.24
FixMatch (RA)**	7.93 ± 1.17	28.02 ± 0.13
GPL-FixMatch** (RA), ours	$\mathbf{6.58} \pm 0.74$	$\mathbf{26.85} \pm 0.48$

Table 1. Error rate of different methods [11, 13, 15, 16, 8] with no assumptions on class distribution on CIFAR-10 with 40 labels and CIFAR-100 with 2500. Values for FixMatch and ReMixMatch taken from [8]. Fully supervised baseline from [11]. *Performance as reported in [14]. **Our runs on 3 folds. CTA stands for the augmentation strategy of [11], and RA for the one of [19].

Imposing that the equilibrium value of L_{GPL} keeps the same magnitude as if $\underline{\lambda} = 0$ and that the two factors of Eq. (5) have the same equilibrium value allows to solve for $\bar{\lambda}$ and $\underline{\lambda}$, so additional hyperparameter fine-tuning is minimal.

There exist portions of the space of images where the derivative of $\mathcal{L}_{GPL}(\cdot; \underline{\lambda} = 0)$ vanishes, but where the derivative of $\mathcal{L}_{GPL}(\cdot; \underline{\lambda})$ with $\underline{\lambda} > 0$ does not. This fact, along with the heuristic supporting better purity for negative than for positive pls completes the intuition behind the better performance of GPL when it replaces PL in any T-S setup, such as PL or FixMatch.

4. EXPERIMENTS: GPL APPLIED ON FIXMATCH

FixMatch is a particular case of our T-S formulation of Eq. (2). Considering a random weak augmentation W and a random strong one S, and calling $f_\theta = B_\theta \circ S$ and $\tilde{g}_\theta = B_\theta \circ W$ where the architecture of the backbone B_θ is fixed, and CE as the loss function, with $PL(\cdot)$ as pseudo-labeling strategy yields the loss function of FixMatch.

GPL-FixMatch is obtained by keeping the same T-S configuration as in FixMatch, replacing PL by GPL. The loss $\bar{\ell}(s)$ on positive pls is $-\log s$ as in FixMatch. The loss on negative pls is $\underline{\ell}(s) = s$, i.e. we use regression on scores for negative pls, as $\underline{\ell}(s) = -\log(1-s)$ led to instabilities.

For all hyperparameters shared with FixMatch, we use the optimal values as obtained in [8], and apply only RA augmentation as in [14]. The values for $\bar{\lambda} = 0.5$ and $\underline{\lambda} = 3.0$ are obtained experimentally from Eq. (5), by doubling the weight $\underline{\lambda}$, based on the intuition that negative pls are more reliable than positive ones. The lower threshold $\underline{\tau}$ is given by the formula of Sec. 3.1. We report the error rate of the last checkpoint of

every experiment on 3 folds using the codebase of [14].

Our experiments on CIFAR-10 and 100 in the regime of scarce labels, 40 and 2500, respectively, showed that GPL-FixMatch improves on FixMatch, with marginal computational overhead and minimal hyperparameter fine-tuning. Since on Imagenet with $100K$ labeled images FixMatch struggles to learn the labeled images throughout training, presenting an accuracy on the labeled dataset of the order of 70%, GPL-FixMatch kept the same (low) level of performance due to insufficient interpolation, which renders the quality of extrapolation irrelevant.

We present the results showing the gains of integrating negative pls in the learning process in Tab. 1. Under the same conditions the improvement of GPL-FixMatch over FixMatch is clear. Introducing a slight class imbalance by dropping 40% of images in a randomly chosen class shows that FixMatch and GPL-FixMatch are more robust than FlexMatch, which heavily capitalizes on the assumption of uniform class distribution, cf. Tab. 2. We also vary the weights in the neighborhood of those obtained experimentally using eq (5), and see that doubling $\underline{\lambda}$ gives better results, see Tab. 3.

FixMatch	FlexMatch	GPL-FixMatch
7.69 ± 1.92	7.88 ± 0.29	$\mathbf{6.56} \pm 0.93$

Table 2. Error rate of FixMatch, FlexMatch and GPL-FixMatch on 3 folds of CIFAR-10-40 with class imbalance.

GPL-FixMatch is also more robust when the hyperparameters of CIFAR-10 are transferred to CIFAR-100 (Tab. 4). In [8] the weight decay was changed from 0.0005 for CIFAR-10 to 0.0001 in an adhoc manner for CIFAR-100. GPL-FixMatch is less affected by the adaptation than FlexMatch. Finally, when we vary the upper threshold (Fig. 2), the accuracy of GPL-FixMatch drops less violently when the upper threshold approaches 1, since it still learns from negative information.

Figure 1 features a comparison between GPL-FixMatch and FixMatch in terms of the accuracy on the test set during training and the proportion of unlabeled images that have been attributed a positive pl. Introduction of GPL clearly results in a better exploitation of unlabeled data. This results in more efficient training, as more information is given to the network, earlier in training, and the learning curve establishes the better quality of this information.

$(\bar{\lambda}, \underline{\lambda})$	$(0.5, 3.0)$	$(1.0, 3.0)$	$(1.0, 6.0)$
error rate	**6.60**	7.28	8.50

Table 3. Error rate of GPL-FixMatch on the first fold of CIFAR-10-40 when the loss weights vary.

hyperparameter adaptation	FixMatch	FlexMatch	GPL FixMatch
No	28.05 $\pm$0.32	27.56 $\pm$0.43	**27.16** $\pm$0.44
Yes	28.02 $\pm$0.13	**26.58** $\pm$0.33	26.85 $\pm$0.48

Table 4. Error rate on 3 folds of CIFAR-100-2500 with weight decay rate transferred from CIFAR-10 to CIFAR-100 (no adaptation) or fine-tuned for CIFAR-100.

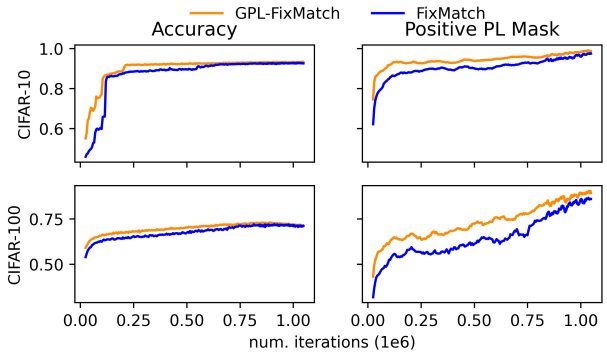

Fig. 1. Learning curves for one fold on CIFAR-10 and CIFAR-100. The model learns better and earlier. The positive PL mask is the proportion of images with a positive pl. Use of negative pls leads to a better data exploitation, both quantitatively (greater number of pls attributed) and qualitatively (accuracy increases).

5. CONCLUSION

We have presented GPL, a simple, efficient and cost-less way to exploit negative information in image classification problems in SSL. We have showcased its efficiency by improving on the baseline of FixMatch with minimal computational overhead and hyperparameter fine-tuning in the regime of scarce labels. Application on semantic segmentation problems (pixel-wise single-class classification) as well as classification problems in NLP could also help improve methods based on PL.

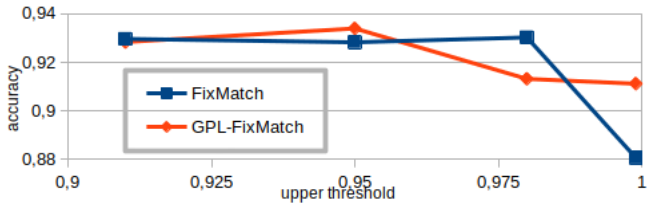

Fig. 2. Accuracy on the first fold of CIFAR-10-40 when the upper threshold varies. GPL makes the performance more robust with respect to the upper threshold.

6. REFERENCES

[1] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020. 1

[2] HM Kabir, Moloud Abdar, Seyed Mohammad Jafar Jalali, Abbas Khosravi, Amir F Atiya, Saeid Nahavandi, and Dipti Srinivasan, "Spinalnet: Deep neural network with gradual input," *arXiv preprint arXiv:2007.03347*, 2020. 1

[3] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021. 1

[4] 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

[5] Yassine Ouali, Céline Hudelot, and Myriam Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020. 1, 2

[6] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, 2020. 1, 2

[7] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, 2013, vol. 3-2, p. 896. 1, 2, 3

[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*, vol. 33, pp. 596–608, 2020. 1, 2, 3, 4

[9] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1, 2

[10] Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, and Quoc V. Le, "Meta pseudo labels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[11] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2019. 1, 2, 3

[12] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8. 1

[13] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin, "Dash: Semi-supervised learning with dynamic thresholding," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11525–11536. 1, 2, 3

[14] 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*, vol. 34, 2021. 1, 2, 3, 4

[15] Byoungjip Kim, Jinho Choo, Yeong-Dae Kwon, Seongho Joe, Seungjai Min, and Youngjune Gwon, "Selfmatch: Combining contrastive self-supervision and consistency for semi-supervised learning," *arXiv preprint arXiv:2101.06480*, 2021. 2, 3

[16] Junnan Li, Caiming Xiong, and Steven Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," *arXiv preprint arXiv:2011.11183*, 2020. 2, 3

[17] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim, "Nlnl: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 101–110. 2

[18] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *International Conference on Learning Representations*, 2020. 2

[19] 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*, 2020, pp. 702–703. 3