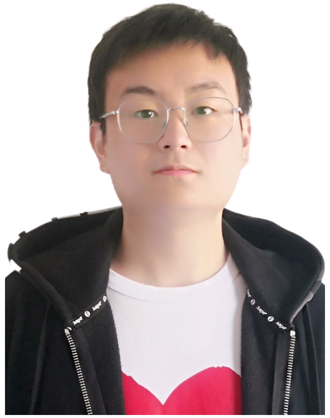
NeurIPS | 2022

Thirty-sixth Conference on Neural Information Processing Systems

Debiased Self-Training for Semi-Supervised Learning

Baixu Chen[*], Junguang Jiang[*], Ximei Wang, Pengfei Wan[§], Jianmin Wang, Mingsheng Long[✉]



Baixu Chen

Junguang Jiang

Ximei Wang

Jianmin Wang

Mingsheng Long

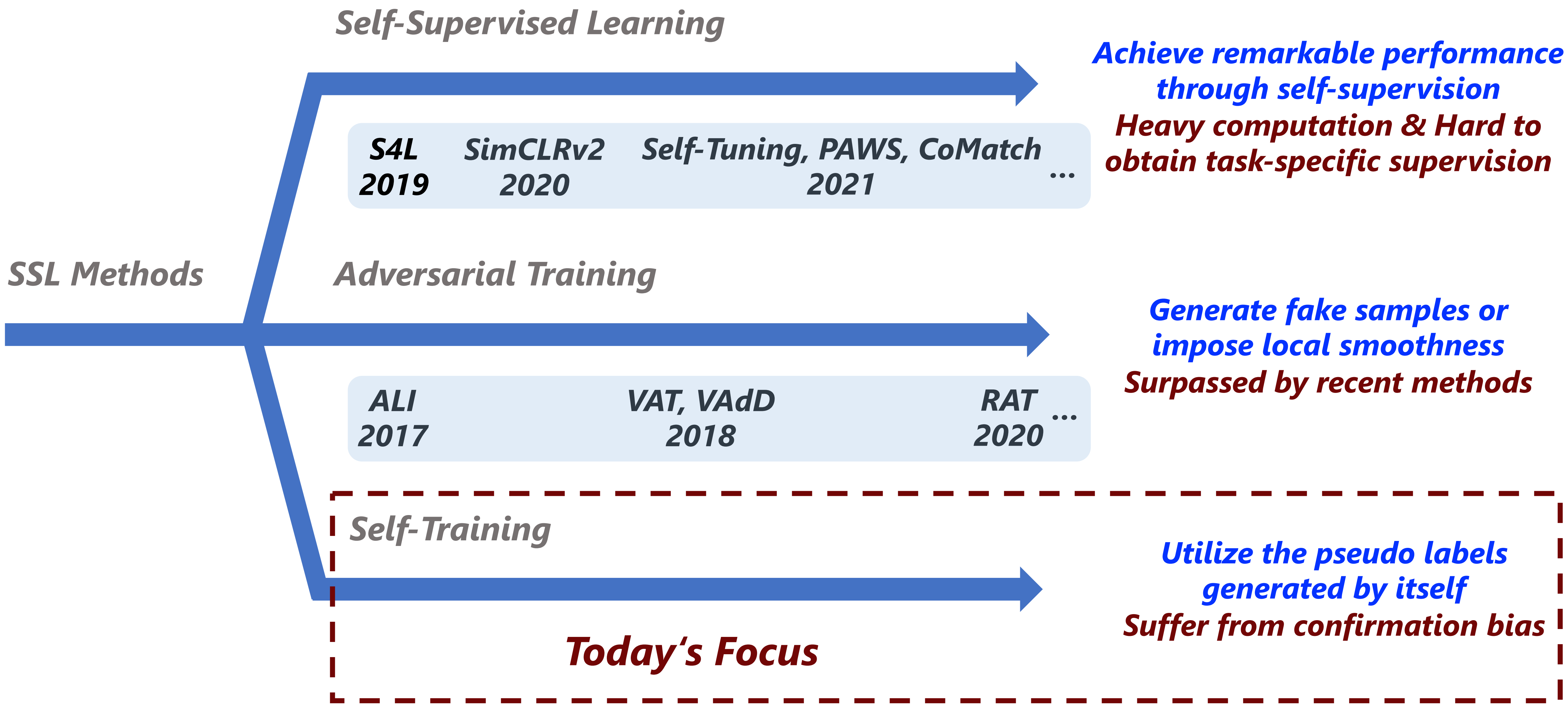
Semi-Supervised Learning (SSL)

- *Aim to improve data efficiency of deep models*
- *Explore supervision from unlabeled data*

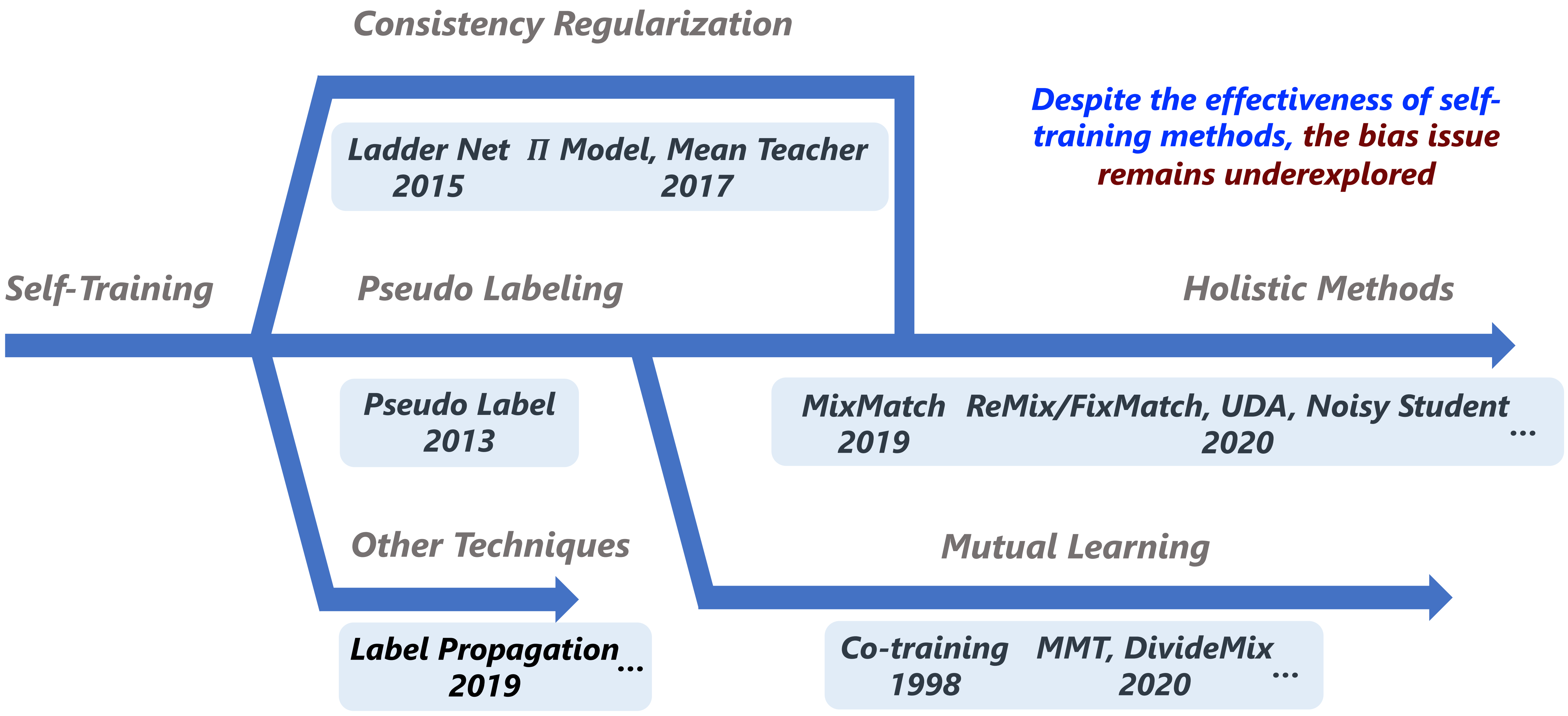
Few Labeled Data $\mathcal{L}$

Numerous Unlabeled Data $\mathcal{U}$

Overview of SSL Methods



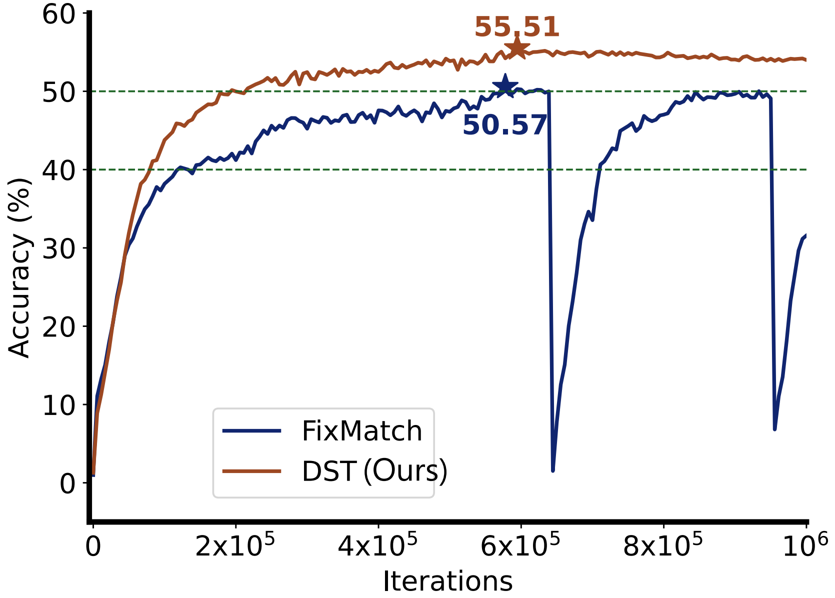
Overview of Self-Training Methods



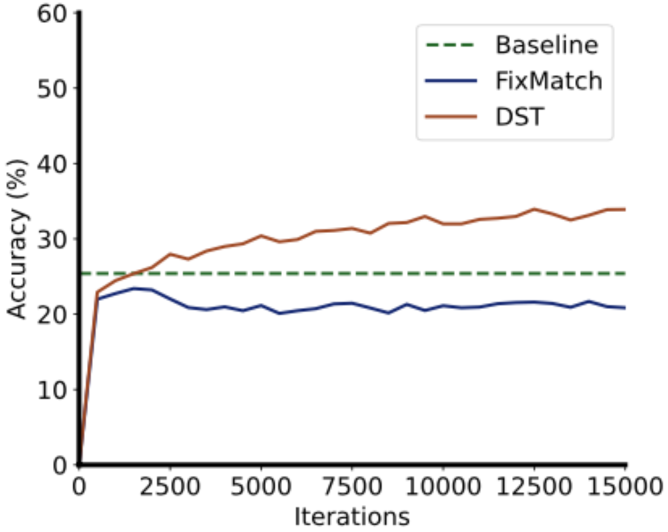
Bias Issue of Self-Training

Training instability

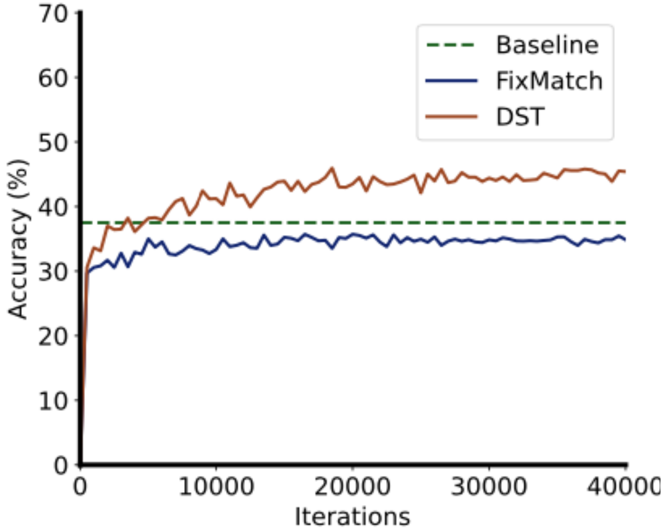
- ***Slow down convergence speed ☹***
- ***Lead to catastrophic forgetting of pre-trained models ☹***



CIFAR-100 (trained from scratch)



Aircraft (supervised pre-trained)

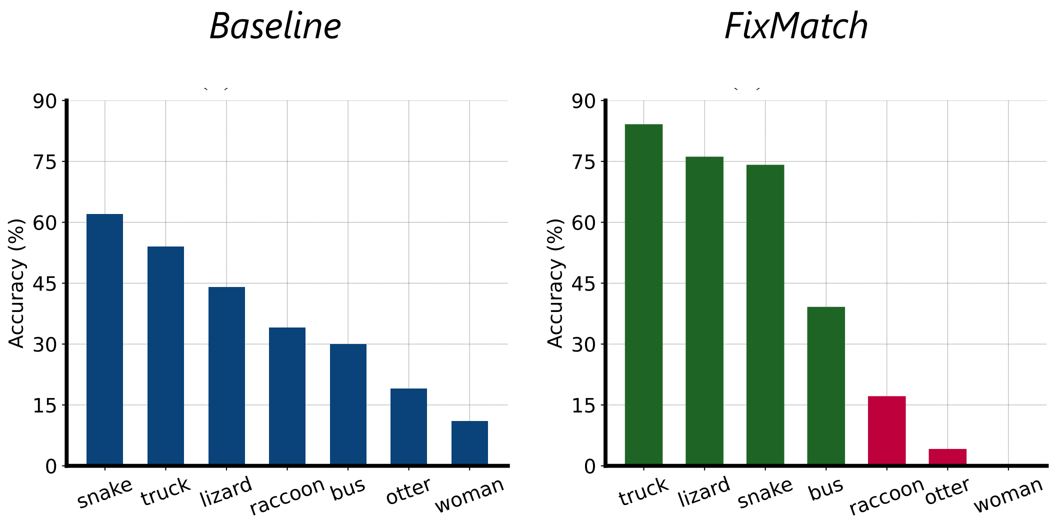


CUB (unsupervised pre-trained)

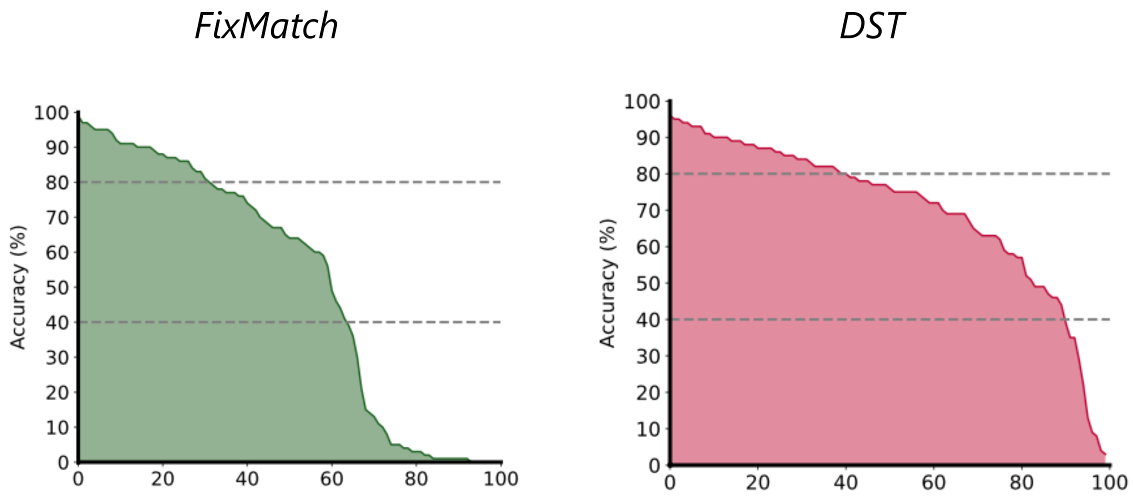
Bias Issue of Self-Training

Matthew Effect

➢ ***Enlarges performance imbalance across classes*** ☹



Top-1 Accuracy on 7 categories from CIFAR-100



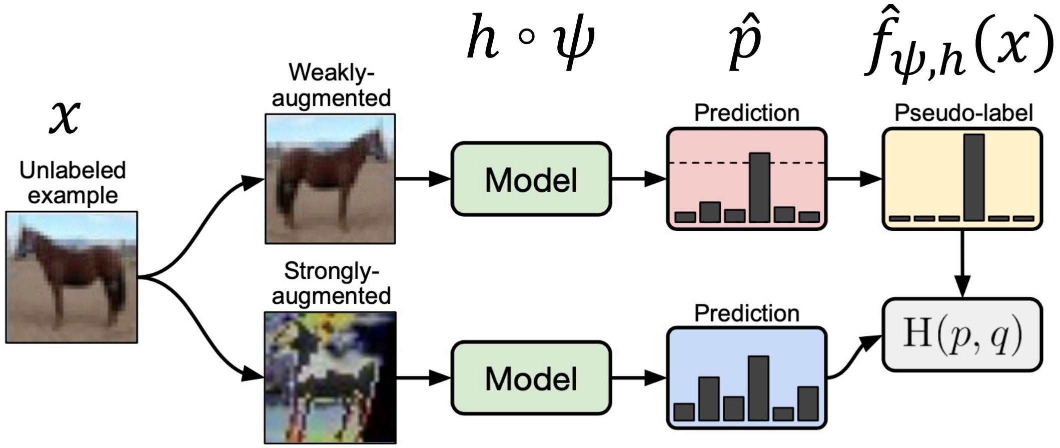
Top-1 Accuracy on all classes from CIFAR-100

Previous Solutions to Self-Training Bias

Generate Higher Quality Pseudo Labels

FixMatch, UDA, FlexMatch …

(1) Confidence Thresholds (Static or Dynamic)
(2) Weak Data Augmentation



Data Flow of FixMatch

$$\hat{f}_{\psi,h}(x) = \begin{cases} \operatorname{argmax} \hat{p}, & \max \hat{p} \geq \tau \\ -1, & \text{otherwise} \end{cases}$$

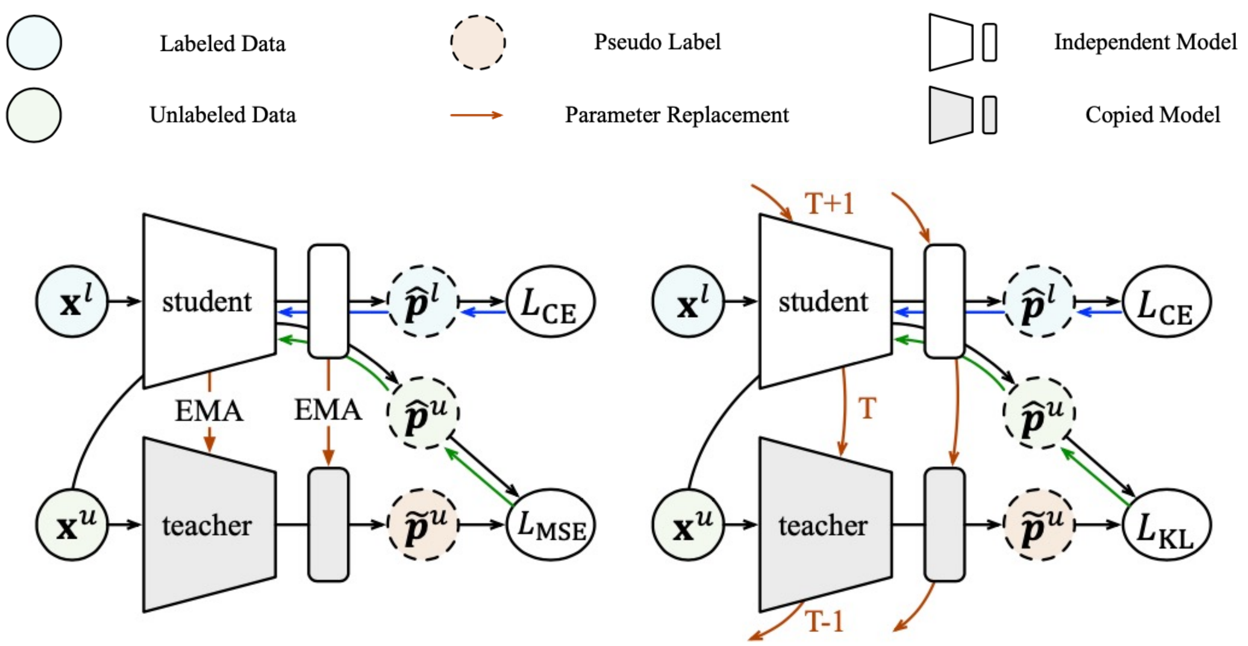
Relies on manual design of criteria to improve the quality of pseudo labels ☹

Previous Solutions to Self-Training Bias

Improve Tolerance to Inaccurate Pseudo Labels

Mean Teacher, MMT, Noisy Student…

Maintain Discrepancy Between Generation and Utilization of Pseudo Labels



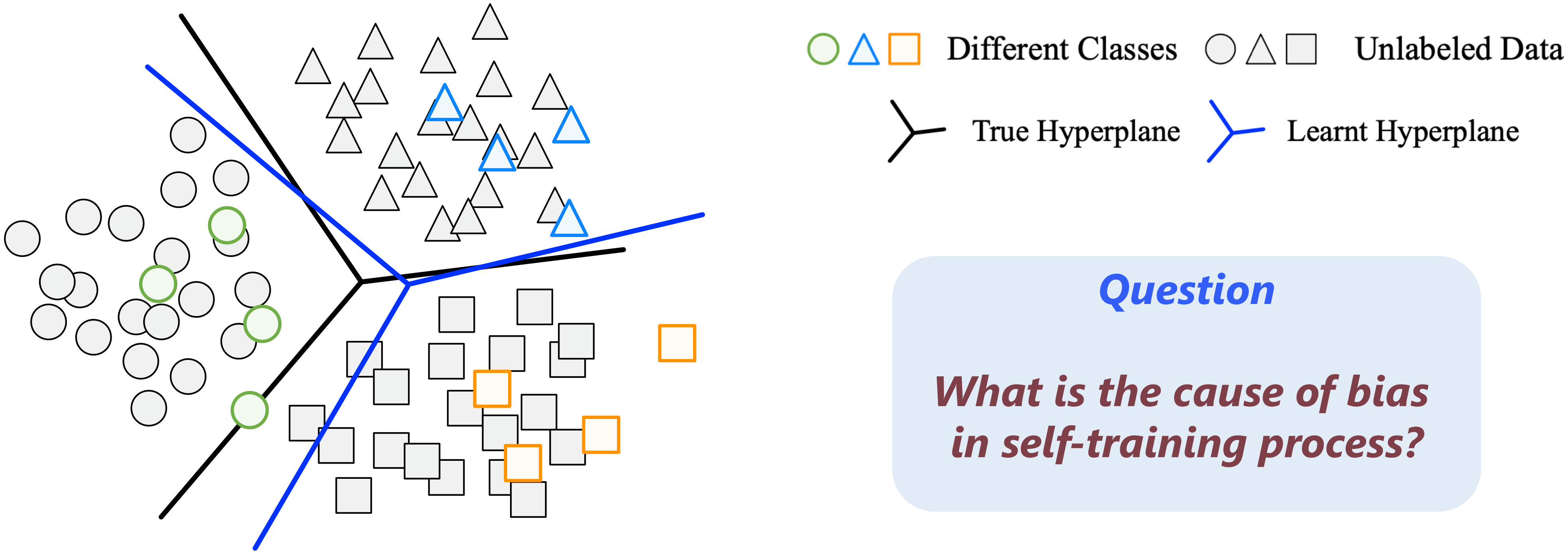
Mean Teacher　　*Noisy Student*

The decision boundary still has the potential to be damaged by incorrect labels ☹

Analysis of Self-Training Bias

Definition of Bias

The deviation between the learned decision hyperplanes and the true decision hyperplanes



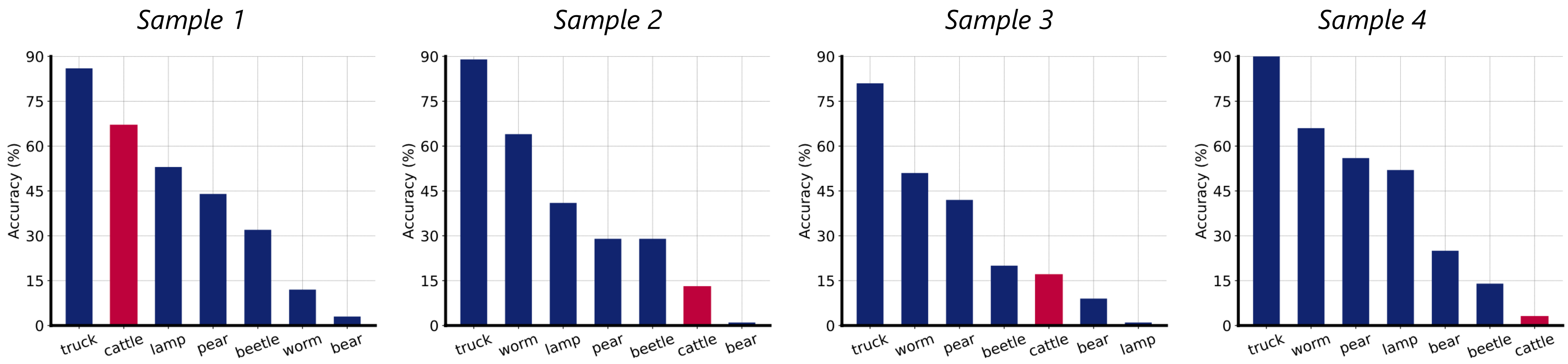
Question

What is the cause of bias in self-training process?

Analysis of Self-Training Bias

Effect of Data Sampling

With fewer data, the distances between supporting data of each category and the true decision hyperplanes may vary

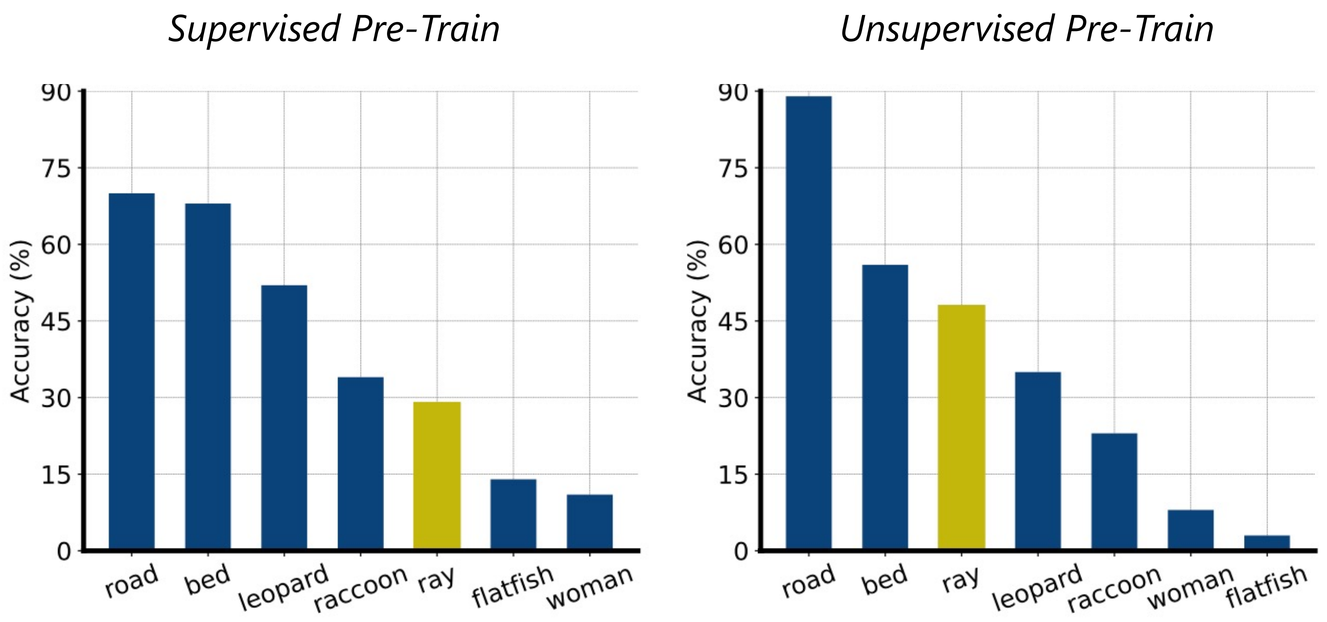


Top-1 Accuracy on 7 categories from CIFAR-100 with different labeled data sampling

Analysis of Self-Training Bias

Effect of Pre-Trained Representations

Different pre-trained models focus on different aspects of the data

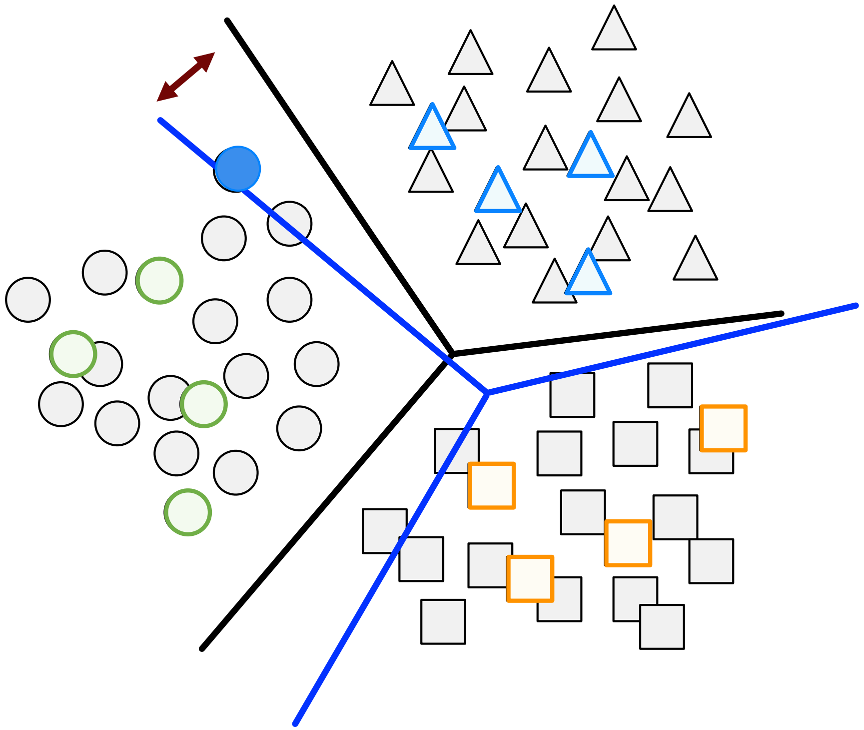


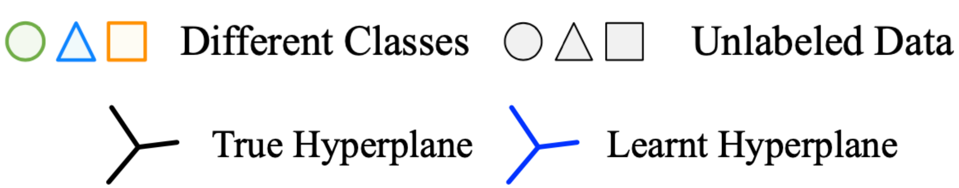
Top-1 Accuracy on 7 categories from CIFAR-100 with different pre-trained models

Analysis of Self-Training Bias

Effect of Self-Training Algorithm

Training with pseudo labels aggressively in turn enlarges the self-training bias on some categories



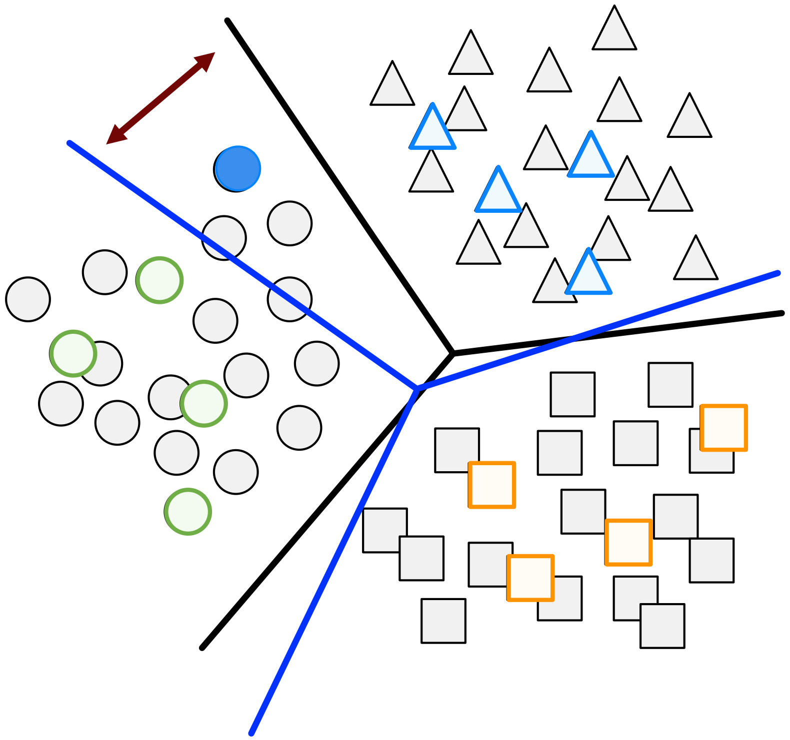


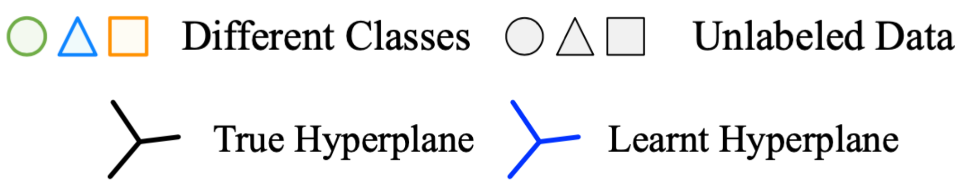
With biased learnt hyperplane, the model makes a mistake on data point

Analysis of Self-Training Bias

Effect of Self-Training Algorithm

Training with pseudo labels aggressively in turn enlarges the self-training bias on some categories



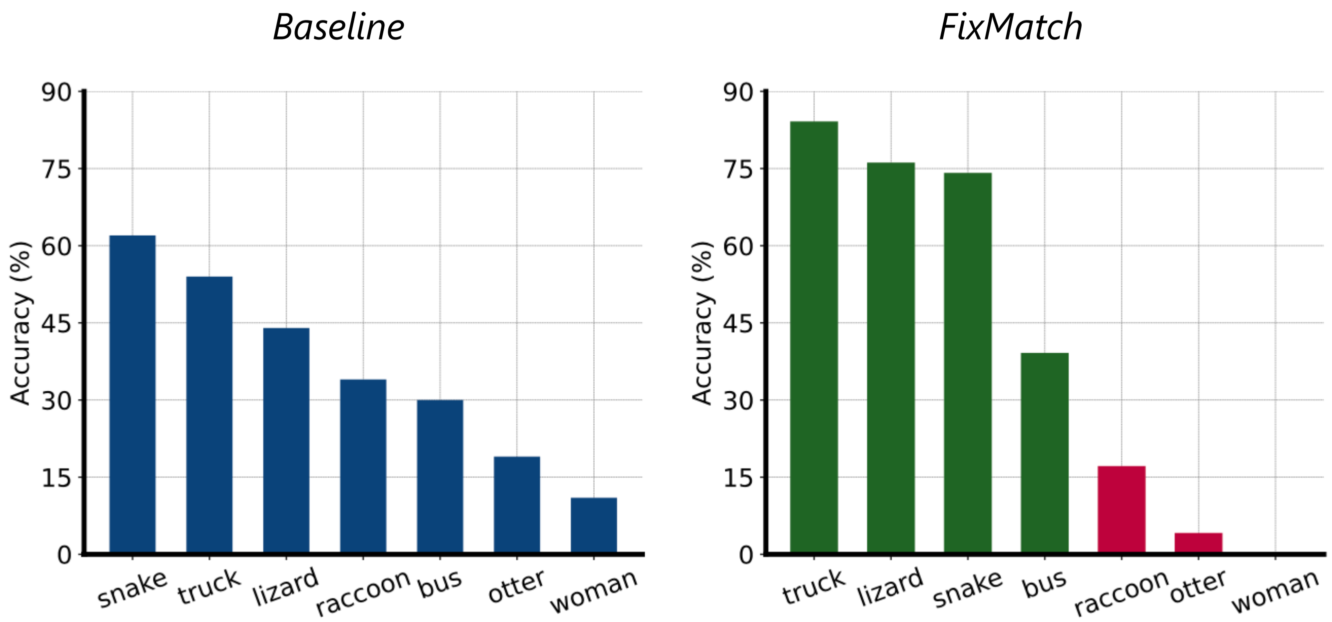


The misclassified data point further pushes the learnt hyperplane far away

Analysis of Self-Training Bias

Effect of Self-Training Algorithm

Ultimately, the accuracy of some categories increases, while that of other categories may decrease to near zero



Top-1 Accuracy on 7 categories from CIFAR-100 with different training strategies

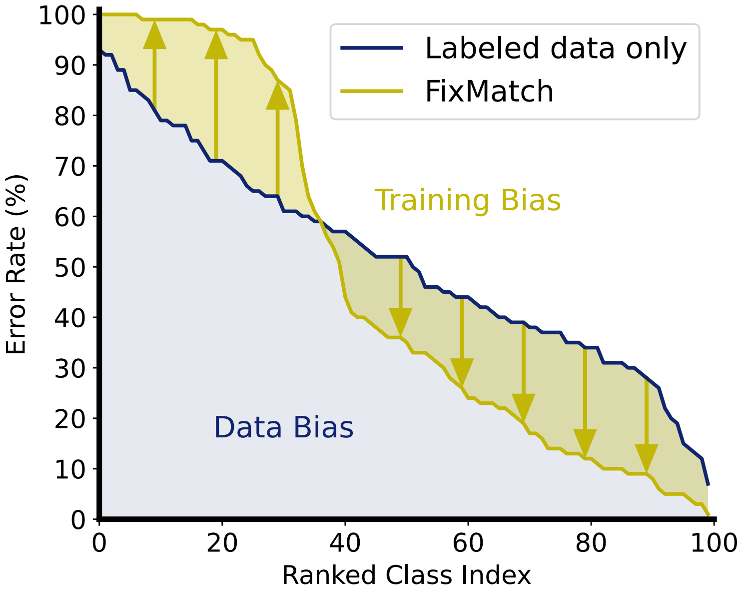
Decomposition of Bias in Self-Training

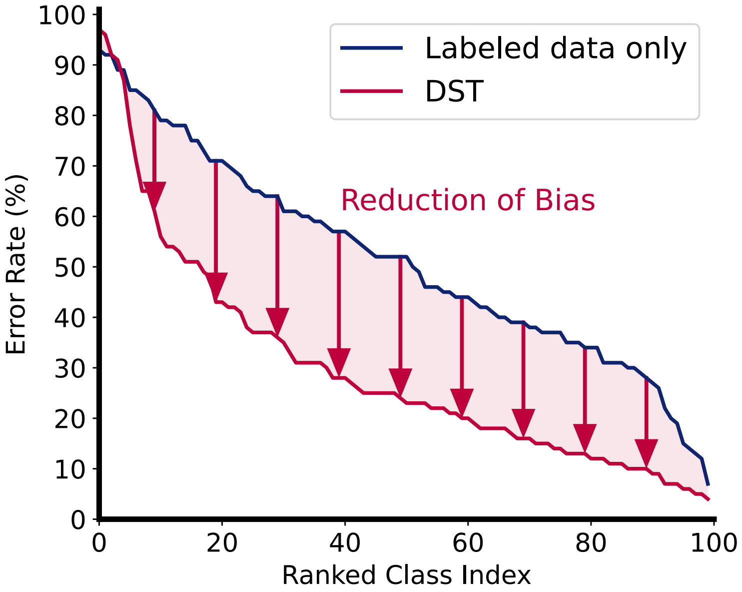
Data Bias

The bias inherent in semi-supervised learning tasks, such as data sampling and pre-trained representations

Training Bias

The bias increment brought by some unreasonable training strategies

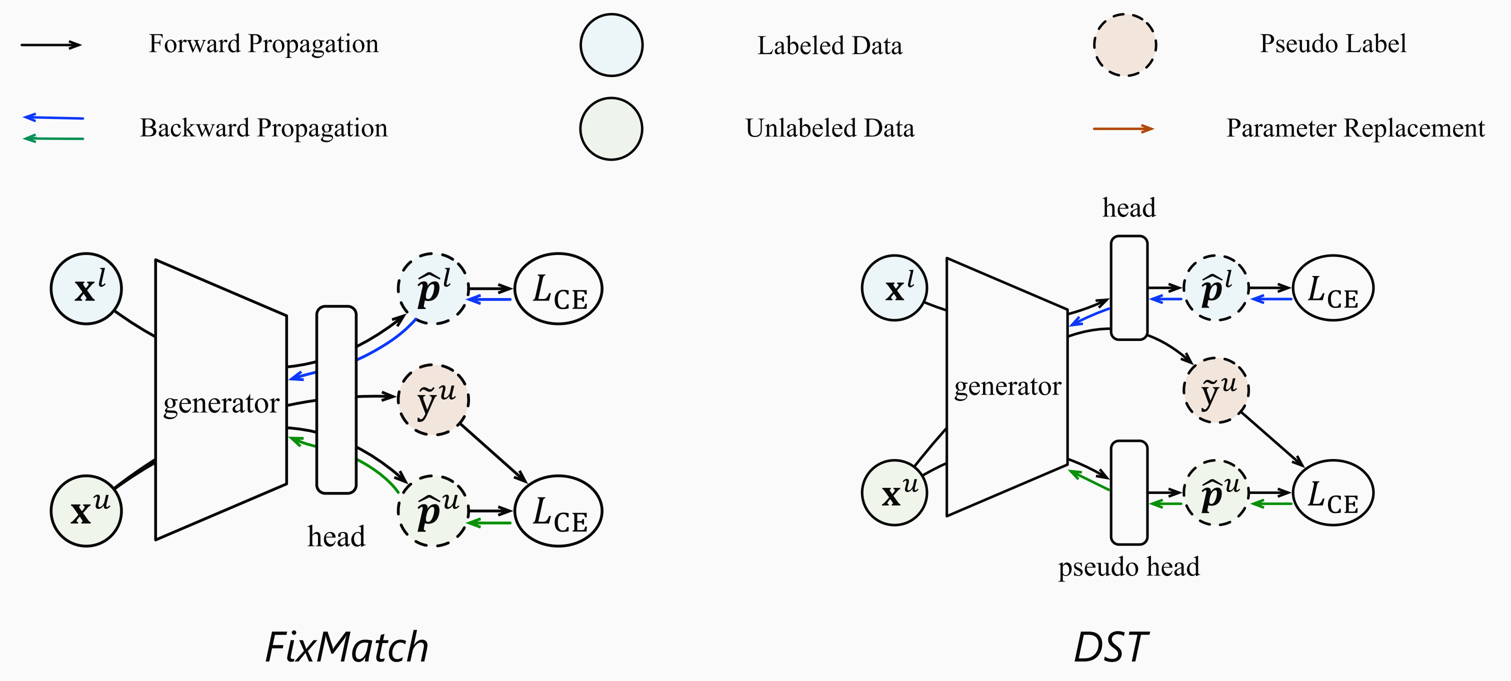




Debiased Self-Training

How to Decrease Training Bias?

***Decouple the generation and utilization of pseudo labels** by introducing a complete parameter-independent pseudo head*

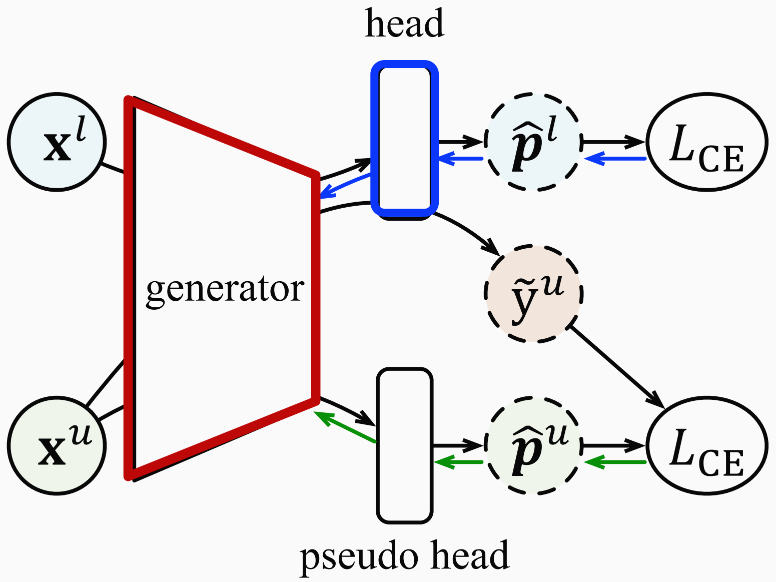


Debiased Self-Training

How to Decrease Training Bias?

$$\min_{\psi, h, h_{\text{pseudo}}} L_{\mathcal{L}}(\psi, h) + \lambda L_{\mathcal{U}}\left(\psi, h_{\text{pseudo}}, \hat{f}_{\psi,h}\right)$$

Insights



Classifier head:

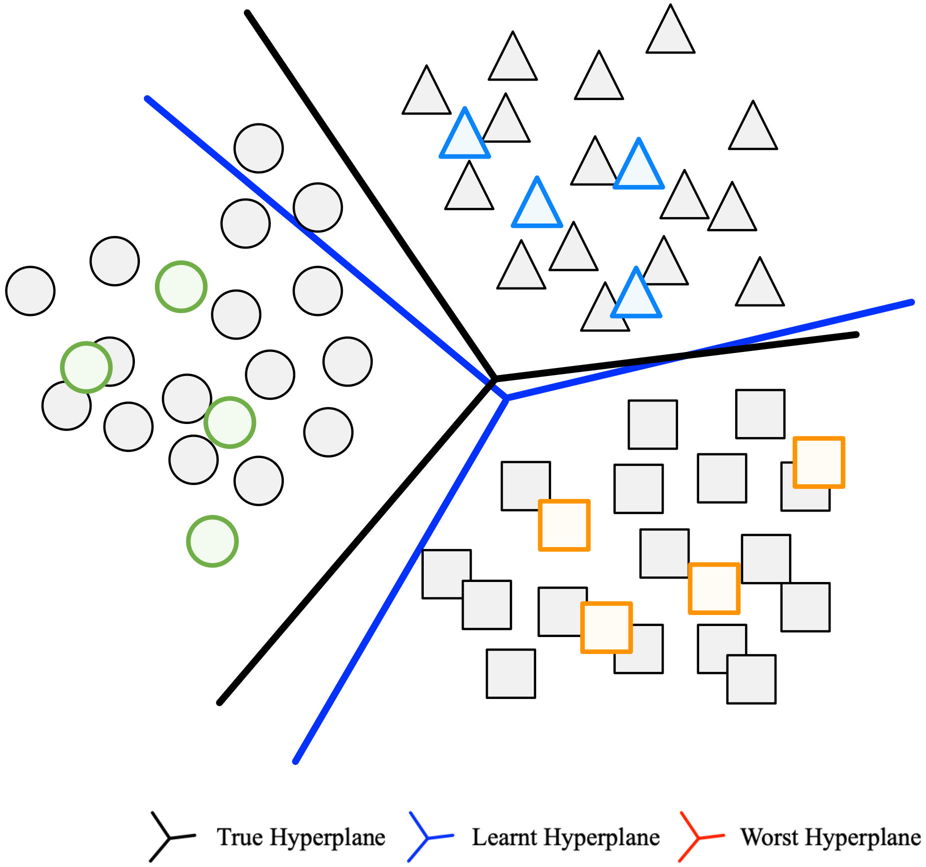
- ***more sensitive to noisy data***

Feature generator:

- ***more parameters, data hungry,***
- ***better tolerance to noisy data***

Debiased Self-Training

How to Decrease Data Bias?

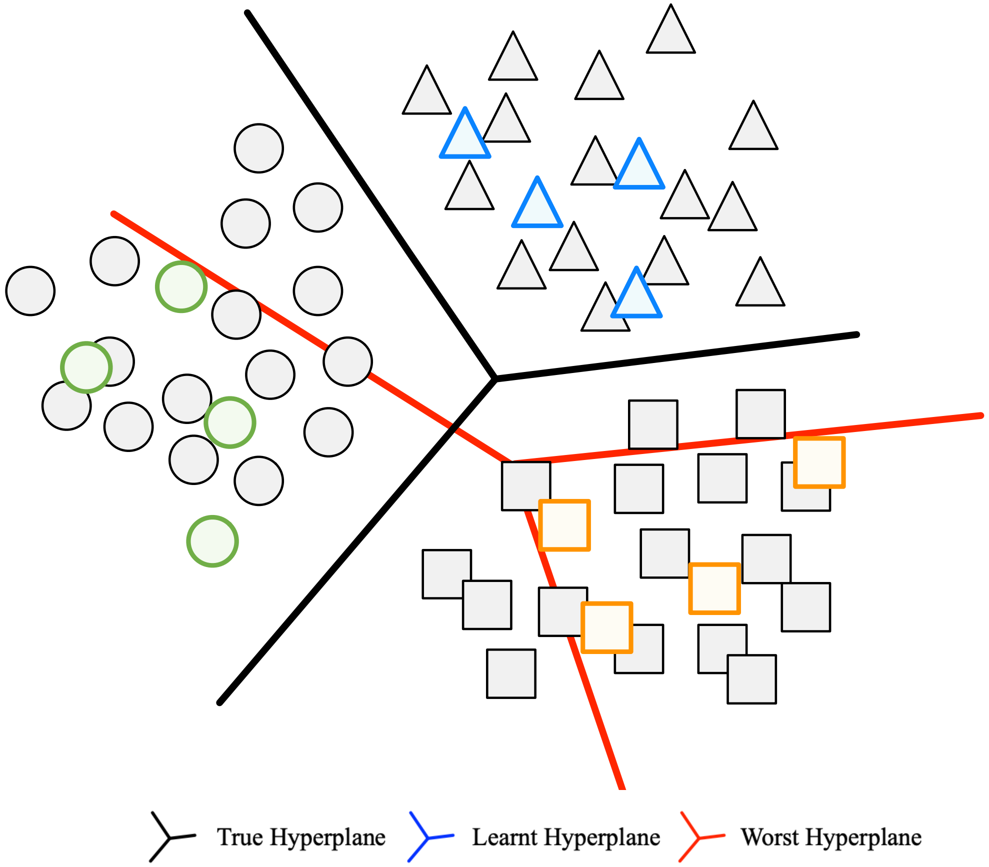


Worst Case Estimation

(1) ***Training bias*** *can be considered as the* ***accumulation of data bias***

Debiased Self-Training

How to Decrease Data Bias?



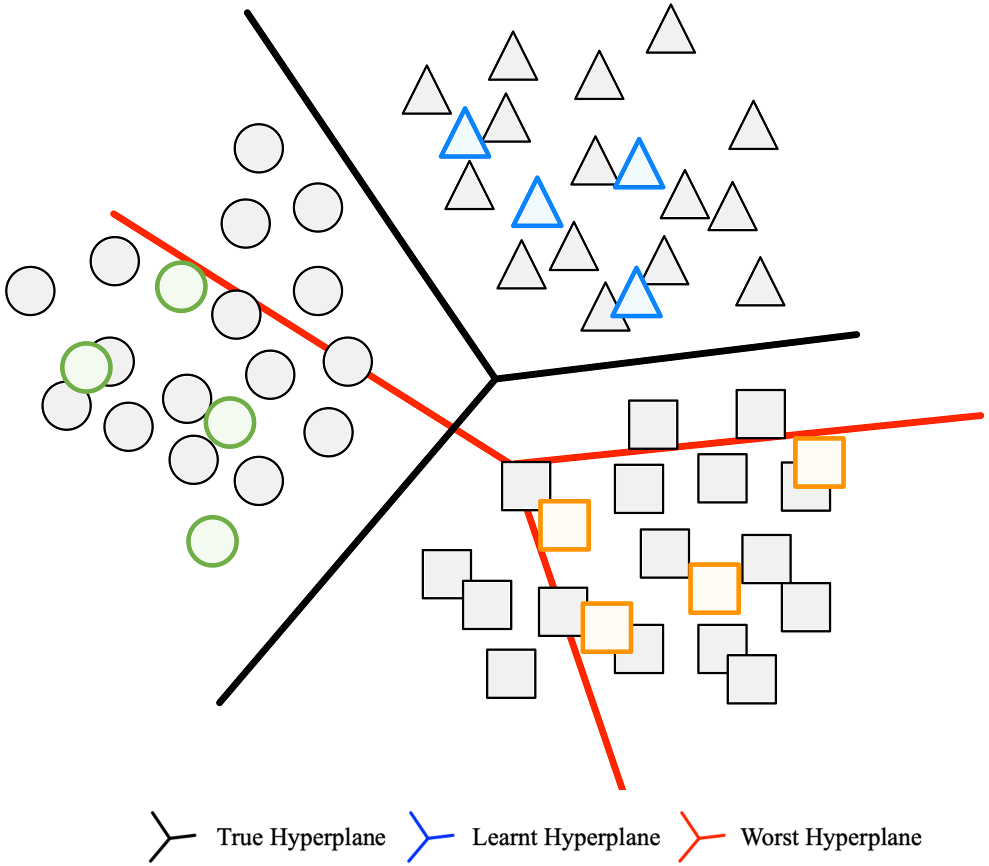
Worst Case Estimation

*(1) **Training bias** can be considered as the **accumulation of data bias***

*(2) The **worst training bias** that can be achieved is a good measure of data bias*

Debiased Self-Training

How to Decrease Data Bias?



Estimate the Worst Training Bias

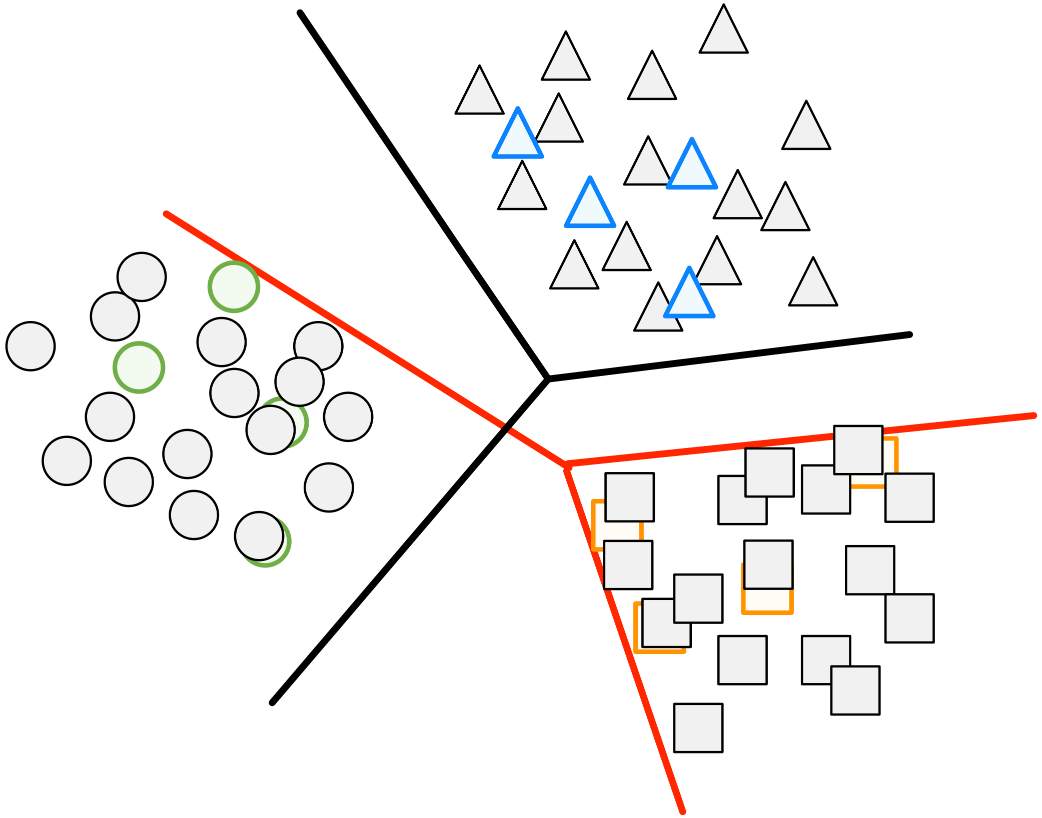
$$\max_{h'} L_{\mathcal{U}}(\psi, h', \hat{f}_{\psi,h}) - L_{\mathcal{L}}(\psi, h')$$

Introduce a worst case estimation head h', which

- ***Correctly classifies the labeled samples***
- ***Deviates from the current hyperplanes as much as possible***

Debiased Self-Training

How to Decrease Data Bias?



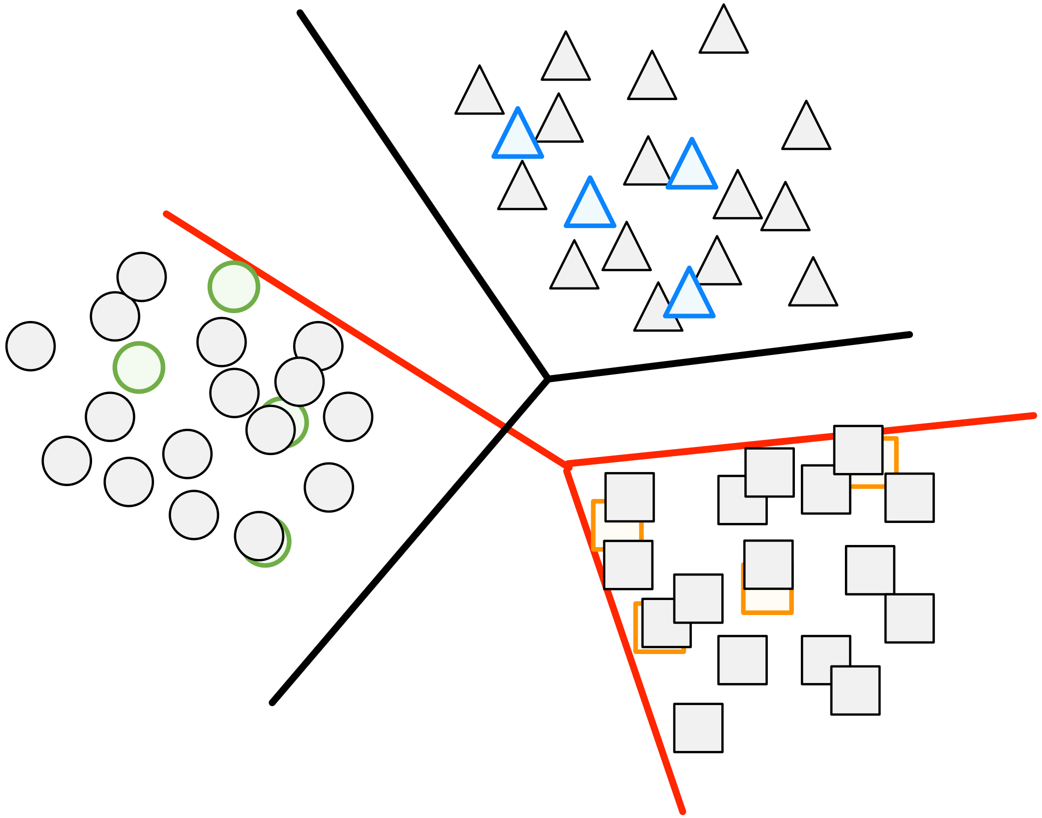
Decrease the Worst Training Bias

$$\min_{\psi} \max_{h'} L_{\mathcal{U}}\left(\psi, h', \hat{f}_{\psi,h}\right) - L_{\mathcal{L}}(\psi, h')$$

Encourage the features to be generated far away from the current hyperplanes

Debiased Self-Training

How to Decrease Data Bias?



Decrease the Worst Training Bias

$$\min_{\psi} \max_{h'} L_{\mathcal{U}}\left(\psi, h', \hat{f}_{\psi,h}\right) - L_{\mathcal{L}}(\psi, h')$$

Implementation

- ***We optimize ψ and h' with stochastic gradient descent alternatively***
- ***The optimization can be viewed as an alternative form of GAN***

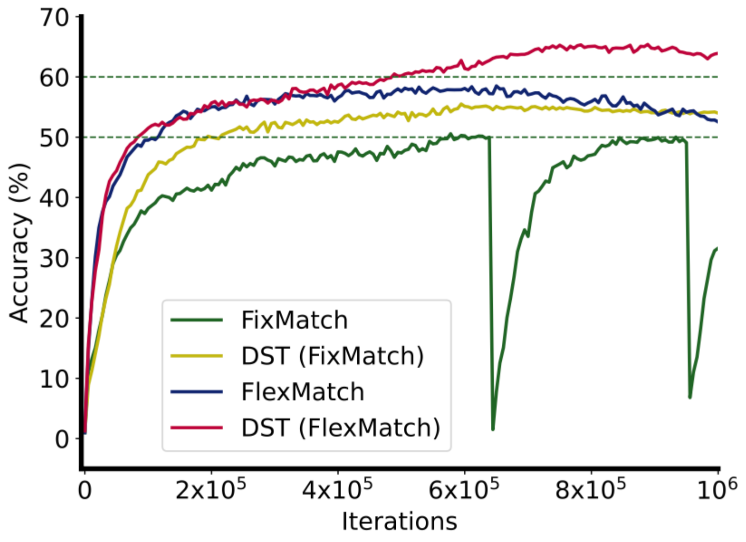
Experiments: Standard SSL Benchmarks

Method	CIFAR-10	CIFAR-100	SVHN	STL-10	Avg
Psuedo Label [30]	25.4	12.6	25.3	25.3	22.2
VAT [34]	25.3	15.1	26.1	25.5	23.0
ALI [15]	25.9	12.4	28.5	24.1	22.7
RAT [52]	33.2	20.5	52.6	30.7	34.2
MixMatch [4]	52.6	32.4	57.5	45.1	46.9
UDA [59]	71.0	40.7	47.4	62.6	55.4
ReMixMatch [3]	80.9	55.7	96.6	64.0	74.3
Dash [61]	86.8	55.2	**97.0**	64.5	75.9
FixMatch [49]	87.2	50.6	96.5	67.1	75.4
DST (FixMatch)	**89.3**	**56.1**	96.7	**71.0**	**78.3**
FlexMatch [64]	94.7	59.5	89.6	71.3	78.8
DST (FlexMatch)	**95.0**	**65.4**	**94.2**	**79.6**	**83.6**

DST achieves new state-of-the-art

Especially on the challenging tasks CIFAR-100 and STL10

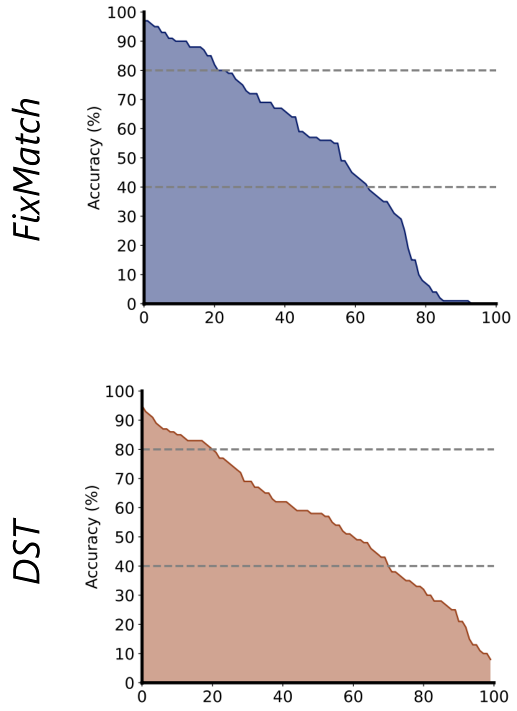
Experiments: Standard SSL Benchmarks



DST helps stabilize training

Method	CIFAR-10	CIFAR-100	SVHN	STL-10	Avg
Pseudo Label [30]	25.4	12.6	25.3	25.3	22.2
VAT [34]	25.3	15.1	26.1	25.5	23.0
ALI [15]	25.9	12.4	28.5	24.1	22.7
RAT [52]	33.2	20.5	52.6	30.7	34.2
MixMatch [4]	52.6	32.4	57.5	45.1	46.9
UDA [59]	71.0	40.7	47.4	62.6	55.4
ReMixMatch [3]	80.9	55.7	96.6	64.0	74.3
Dash [61]	86.8	55.2	**97.0**	64.5	75.9
FixMatch [49]	87.2	50.6	96.5	67.1	75.4
DST (FixMatch)	**89.3**	**56.1**	96.7	**71.0**	**78.3**
FlexMatch [64]	94.7	59.5	89.6	71.3	78.8
DST (FlexMatch)	**95.0**	**65.4**	**94.2**	**79.6**	**83.6**

Experiments: Standard SSL Benchmarks



DST improves performance balance

Method	CIFAR-10	CIFAR-100	SVHN	STL-10	Avg
Psuedo Label [30]	25.4	12.6	25.3	25.3	22.2
VAT [34]	25.3	15.1	26.1	25.5	23.0
ALI [15]	25.9	12.4	28.5	24.1	22.7
RAT [52]	33.2	20.5	52.6	30.7	34.2
MixMatch [4]	52.6	32.4	57.5	45.1	46.9
UDA [59]	71.0	40.7	47.4	62.6	55.4
ReMixMatch [3]	80.9	55.7	96.6	64.0	74.3
Dash [61]	86.8	55.2	**97.0**	64.5	75.9
FixMatch [49]	87.2	50.6	96.5	67.1	75.4
DST (FixMatch)	**89.3**	**56.1**	96.7	**71.0**	**78.3**
FlexMatch [64]	94.7	59.5	89.6	71.3	78.8
DST (FlexMatch)	**95.0**	**65.4**	**94.2**	**79.6**	**83.6**

Experiments: SSL with Supervised Pre-trained Models

		Caltech101	CIFAR-10	CIFAR-100	SUN397	DTD	Aircraft	CUB	Flowers	Pets	Cars	Food101	Average
Supervised	Baseline	81.4	65.2	48.2	39.9	47.7	25.4	46.5	85.2	78.1	33.3	33.8	53.2
	Pseudo Label [30]	86.3	83.3	54.7	41.0	50.2	27.2	54.3	92.3	87.8	41.4	38.0	59.7
	Π-Model [29]	83.5	73.1	49.2	39.7↓	50.3	24.3↓	47.1	90.7	82.2	30.9↓	33.9	55.0
	Mean Teacher [53]	83.7	82.1	56.0	37.9↓	51.6	30.7	49.6	91.0	82.8	39.1	40.3	58.6
	VAT [34]	84.1	72.2	48.8	39.5↓	50.6	25.9	48.1	89.4	81.8	32.4↓	36.7	55.4
	ALI [15]	82.2	69.5	46.3↓	36.4↓	50.5	21.3↓	42.5↓	82.9↓	77.4↓	29.8↓	31.7↓	51.9
	RAT [52]	84.0	81.8	55.4	39.0↓	49.1	31.6	50.0	89.9	84.1	37.9	38.4	58.3
	MixMatch [4]	85.4	82.8	53.5	41.8	50.1	24.7↓	51.7	91.5	83.3	42.5	38.2	58.7
	UDA [59]	85.8	83.6	54.7	41.3	49.0	27.1	52.1	92.0	83.1	45.6	41.7	59.6
	FixMatch [49]	86.3	84.6	53.1	41.3	48.6	25.2↓	52.3	93.2	83.7	46.4	37.1	59.3
	Self-Tuning [55]	87.2	76.0	57.1	41.8	50.7	35.2	58.9	92.6	86.6	58.3	41.9	62.4
	FlexMatch [64]	87.1	89.0	63.4	48.3	52.5	34.0	54.9	94.5	88.3	57.5	49.5	65.4
	DebiasMatch [56]	88.6	91.0	65.7	46.6	52.4	37.5	58.6	95.6	86.4	60.5	53.5	66.9
	DST (FixMatch)	89.6	94.9	70.4	48.1	53.5	43.2	68.7	94.8	89.8	71.0	**58.5**	71.1
	DST (FlexMatch)	**90.6**	**95.9**	**71.2**	**49.8**	**56.2**	**44.5**	**70.5**	**95.8**	**90.4**	**72.7**	57.1	**72.2**

DST achieves the best performance on all datasets

Experiments: SSL with Unsupervised Pre-trained Models

Unsupervised	Baseline	79.5	66.6	46.5	38.1	47.9	28.7	37.5	87.7	60.0	38.1	32.9	51.2
	Pseudo Label [30]	86.2	70.8	49.8	38.6	50.0	26.6↓	41.8	93.0	68.4	37.3↓	32.8↓	54.1
	Π-Model [29]	80.1	76.2	44.8↓	37.8↓	50.0	23.5↓	31.6↓	93.1	62.8	25.6↓	30.4↓	50.5
	Mean Teacher [53]	80.4	80.8	51.3	34.2↓	48.8	33.8	41.6	92.9	67.0	50.5	39.1	56.4
	VAT [34]	79.9	73.8	45.1↓	38.3	49.2	24.2↓	36.4↓	92.4	61.7	29.9↓	33.1	51.3
	ALI [15]	76.4↓	69.2	44.4↓	34.9↓	50.1	22.2↓	33.8↓	84.9↓	59.6↓	33.1↓	31.0↓	49.1
	RAT [52]	80.9	79.5	52.4	37.0↓	50.4	30.1	40.7	91.8	70.5	47.9	35.6	56.1
	MixMatch [4]	84.1	81.5	51.7	38.4	47.0↓	31.7	39.8	93.5	66.4	47.1	34.6	56.0
	UDA [59]	85.0	87.4	53.6	42.3	46.2↓	35.7	41.4	94.1	69.3	51.5	39.3	58.7
	FixMatch [49]	83.1	82.2	51.4	39.2	43.9↓	30.1	36.8↓	94.3	65.7	48.6	36.8	55.6
	Self-Tuning [55]	81.6	63.6↓	47.8	38.8	45.5↓	31.4	41.6	91.0	66.9	52.0	34.0	54.0
	FlexMatch [64]	86.4	96.7	60.2	45.3	53.9	42.0	49.2	95.8	72.9	69.0	37.5	64.4
	DebiasMatch [56]	86.4	96.3	66.3	44.5	53.9	44.8	51.2	95.4	70.9	72.5	53.6	66.9
	DST (FixMatch)	90.1	95.0	68.2	46.8	54.2	**47.7**	53.6	95.6	**75.4**	72.0	**57.1**	68.7
	DST (FlexMatch)	**90.4**	**96.9**	**68.9**	**48.8**	**55.9**	47.3	**55.2**	**96.4**	75.1	**74.6**	56.9	**69.7**

Again, DST achieves the best performance on all datasets

On average, DST surpasses FixMatch by over 15%

Experiments: Ablation Study

Method	Multiple Heads	Linear Pseudo Head	Nonlinear Pseudo Head	Worst Case Estimation	Supervised Pre-training	Unsupervised Pre-training
FixMatch					53.1	51.4
Mutual Learning	✓				53.4	52.5
DST w/o worst	✓	✓			58.2	59.0
DST w/o worst	✓		✓		60.6	60.9
DST	✓		✓	✓	**70.4**	**68.2**

(1) Compared with Mutual Learning, the decoupled pseudo labeling in DST can better reduce training bias

Experiments: Ablation Study

Method	Multiple Heads	Linear Pseudo Head	Nonlinear Pseudo Head	Worst Case Estimation	Supervised Pre-training	Unsupervised Pre-training
FixMatch					53.1	51.4
Mutual Learning	✓				53.4	52.5
DST w/o worst	✓	✓			58.2	59.0
DST w/o worst	✓		✓		60.6	60.9
DST	✓		✓	✓	**70.4**	**68.2**

(1) Compared with Mutual Learning, the decoupled pseudo labeling in DST can better reduce training bias

(2) A nonlinear pseudo head is always better than a linear pseudo one. Possibly because it can reduce the degeneration of representation with biased pseudo labels

Experiments: Ablation Study

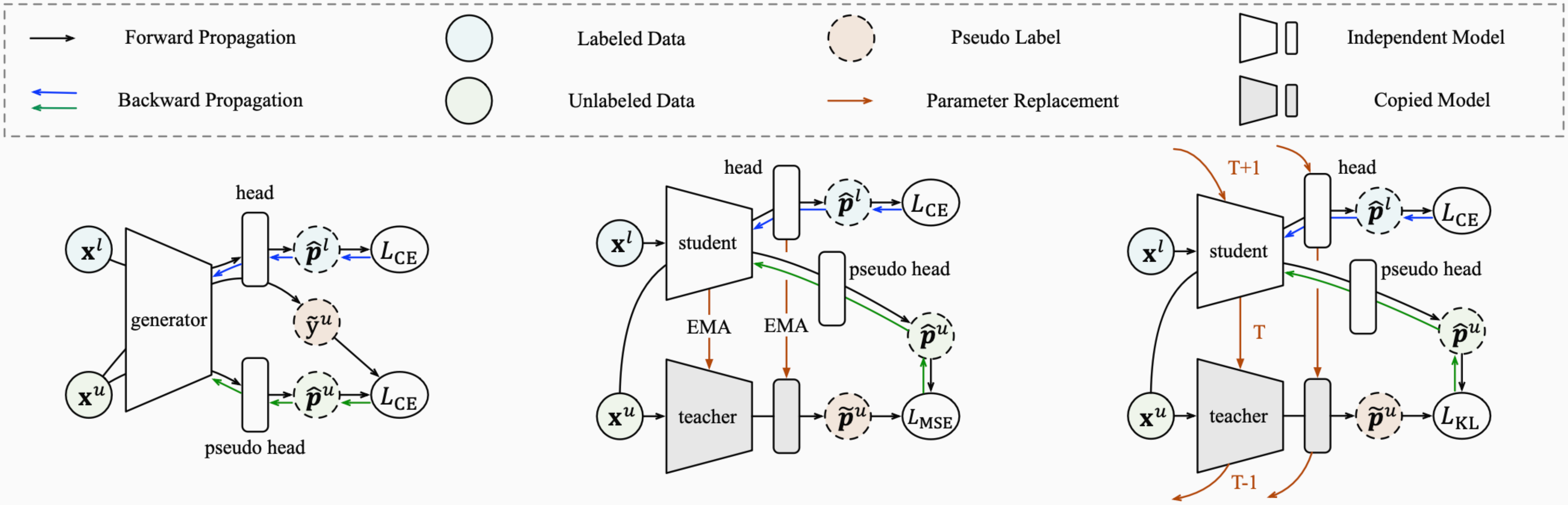
Method	Multiple Heads	Linear Pseudo Head	Nonlinear Pseudo Head	Worst Case Estimation	Supervised Pre-training	Unsupervised Pre-training
FixMatch					53.1	51.4
Mutual Learning	✓				53.4	52.5
DST w/o worst	✓	✓			58.2	59.0
DST w/o worst	✓		✓		60.6	60.9
DST	✓		✓	✓	**70.4**	**68.2**

(1) Compared with Mutual Learning, the decoupled pseudo labeling in DST can better reduce training bias

(2) A nonlinear pseudo head is always better than a linear pseudo one. Possibly because it can reduce the degeneration of representation with biased pseudo labels

(3) The worst-case estimation of pseudo labeling improves the performance by large margins

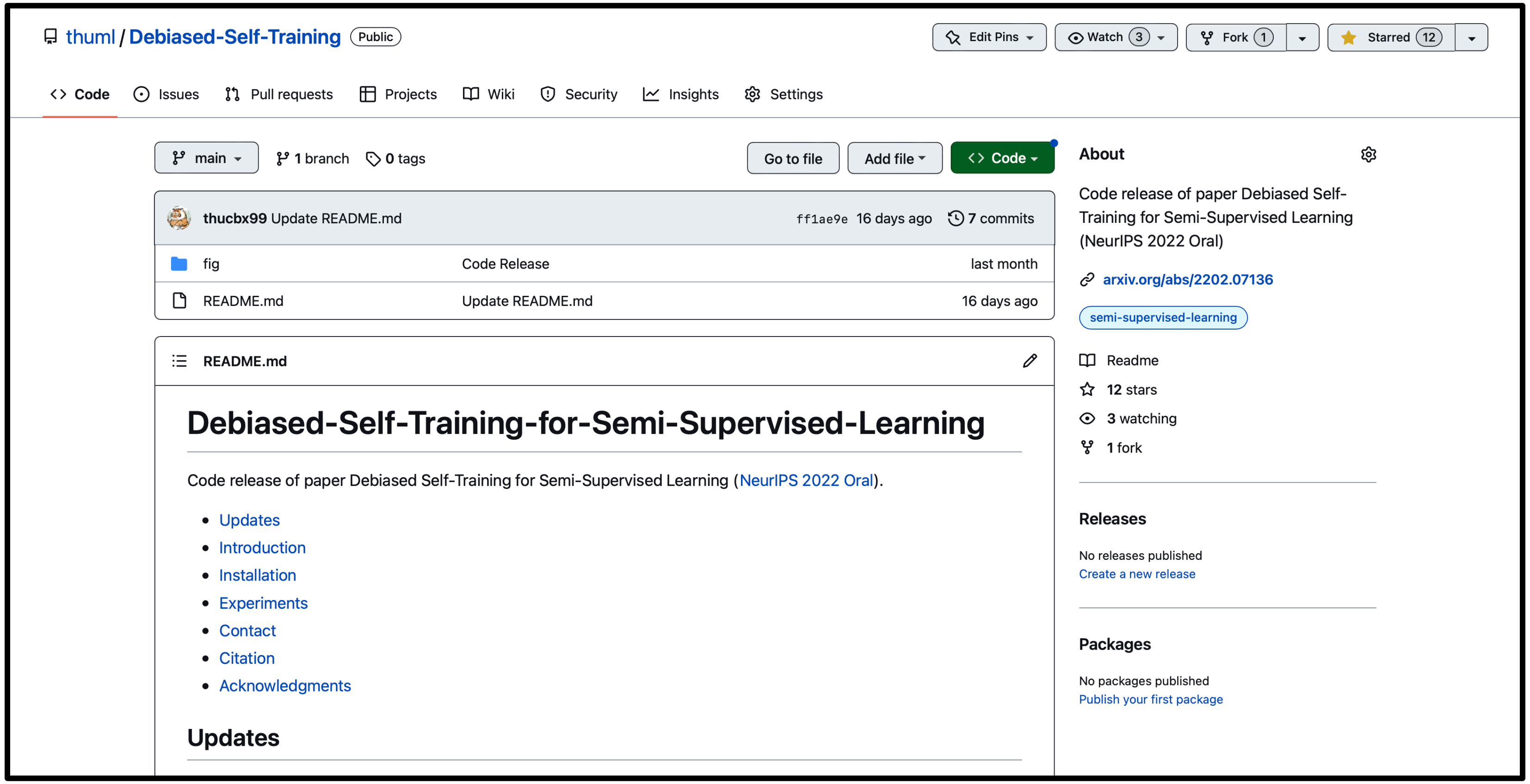
Experiments: DST as a General Add-on



Debiased FixMatch / FlexMatch   *Debiased Mean Teacher*   *Debiased Noisy Student*

DST can be seamlessly incorporated into mainstream self-training methods to reduce bias and boost their performance

Open Source



https://github.com/thuml/Debiased-Self-Training

Complete benchmarks & datasets & scripts

Thank You!

cbx22@mails.tsinghua.edu.cn
jjg20@mails.tsinghua.edu.cn

长按关注，获取最新资讯