

Meta Pseudo Labels

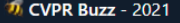
Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, Quoc V. Le

Google AI, Brain Team

CVPR 2021

Presenter: Minho Park

CVPR Buzz – 2021 (Top 1)

 CVPR Buzz - 2021

Sorting Weights

Citations

Replies

Retweets

Likes

1.00

0.25

0.50

0.25

Sort

Poster Session

☐ Monday

☐ Tuesday

☐ Wednesday

☐ Thursday

☐ Friday

Abstracts

Full

Preview



Hide

CVPR Buzz - 2021

Built by [Matt Deitke](#)

CVPR Buzz displays the most discussed papers at CVPR 2021 using Twitter for indexing discussions and Semantic Scholar for collecting citation data.

To add data or see how it was collected, checkout the GitHub repo:



 /mattdeitke/cvpr-buzz  0

1660 results

[1] Meta Pseudo Labels

Hieu Pham, Zihang Dai, Qizhe Xie, Quoc V. Le

We present Meta Pseudo Labels, a semi-supervised learning method that achieves a new state-of-the-art top-1 accuracy of 90.2% on ImageNet, which is 1.6% better than the existing state-of-the-art. [\[Expand\]](#)

 PDF  Semantic Scholar  arXiv  Show Tweets

 1178.75  36  44  723  3081

 Thursday Poster Session


[2] Animating Pictures With Eulerian Motion Fields

Aleksander Holynski, Brian L. Curless, Steven M. Seitz, Richard Szeliski

In this paper, we demonstrate a fully automatic method for converting a still image into a realistic animated looping video. [\[Expand\]](#)

 PDF  Semantic Scholar  arXiv  Show Tweets

 1114.50  2  130  686  2948

 Tuesday Poster Session

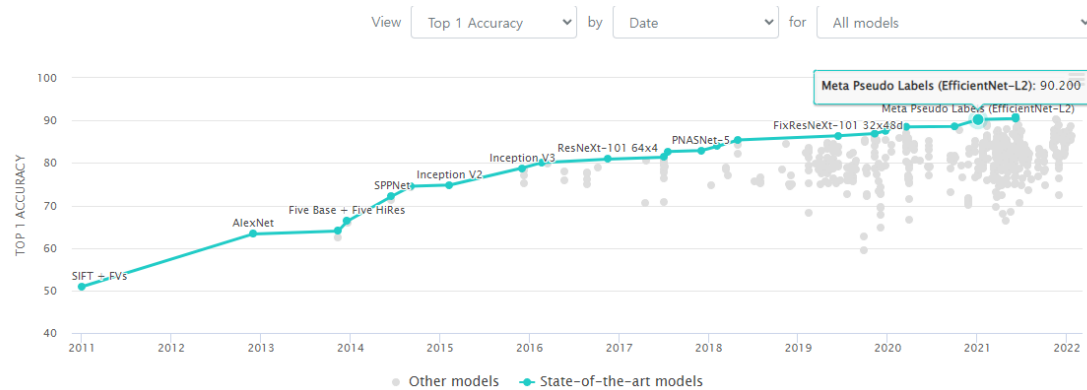
[3] Taming Transformers for High-Resolution Image Synthesis

Patrick Esser, Robin Rombach, Bjorn Ommer

Designed to learn long-range interactions on sequential data, transformers continue to show state-of-the-art results on a wide variety of tasks. [\[Expand\]](#)

ImageNet-1K Classification

- It was the state-of-the-art model on ImageNet-1K and still ranks 5.
- It is a **training strategy** that can be simply applied to the other SOTA architectures.



Datasets	ImageNet Top-1 Accuracy	ImageNet-Real Precision@1
Previous SOTA [16, 14]	88.6	90.72
Ours	90.2	91.02

Table 1: Summary of our key results on ImageNet ILSVRC 2012 validation set [56] and the ImageNet-Real test set [6].

Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params	Extra Training Data	Paper	Code	Result	Year	Tags
1	CoAtNet-7	90.88%		2440M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes	GitHub	Result	2021	Conv+Transformer JFT-3B
2	ViT-G/14	90.45%		1843M	✓	Scaling Vision Transformers	GitHub	Result	2021	Transformer JFT-3B
3	CoAtNet-6	90.45%		1470M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes	GitHub	Result	2021	Conv+Transformer JFT-3B
4	ViT-MoE-15B (Every-2)	90.35%		14700M	✓	Scaling Vision with Sparse Mixture of Experts	GitHub	Result	2021	Transformer JFT-3B
5	Meta Pseudo Labels (EfficientNet-L2)	90.2%	98.8%	480M	✓	Meta Pseudo Labels	GitHub	Result	2021	EfficientNet JFT-300M

captured on 6 Feb. 2022

Pseudo Labels

- Two models: teacher and student.
- The teacher model generates pseudo labels on the unlabeled dataset (trained on the labeled dataset).
- The student model trains on both the labeled and the pseudo labeled datasets.

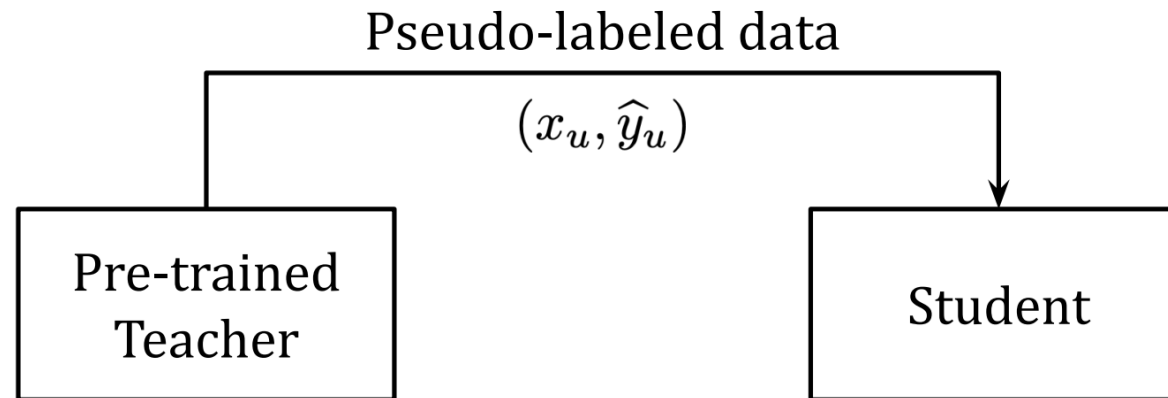


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels (Left).

Pseudo Labels

- The student learns to become better than the teacher.
 - Because of the abundance of pseudo labeled data and the regularization methods.
- **Main drawback:**
 - The pseudo labels can be inaccurate.
 - **The student may not get significantly better than the teacher.**

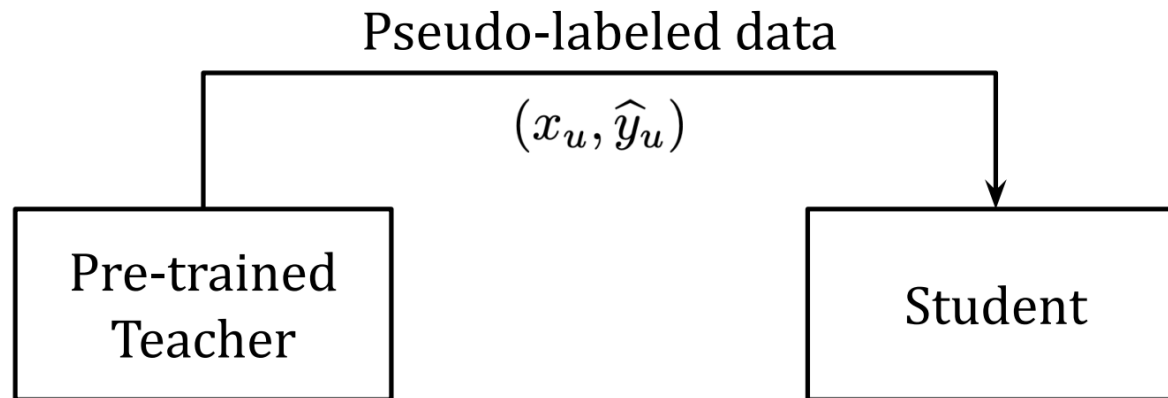


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels (Left).

Meta Pseudo Labels

- A systematic mechanism for the teacher to correct the bias by observing how its pseudo labels would affect the student.

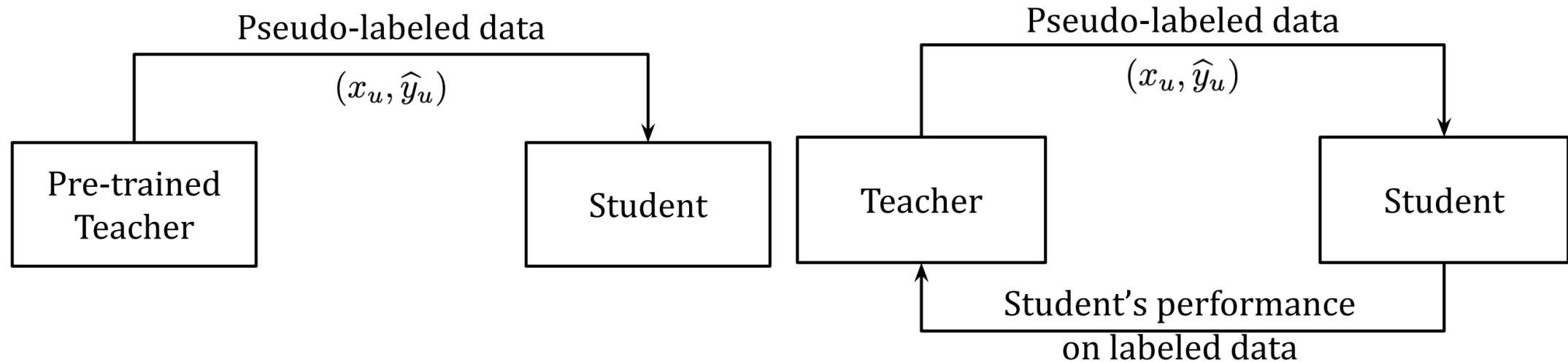


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels.

Meta Pseudo Labels

- A systematic mechanism for the teacher to correct the bias by observing how its pseudo labels would affect the student.

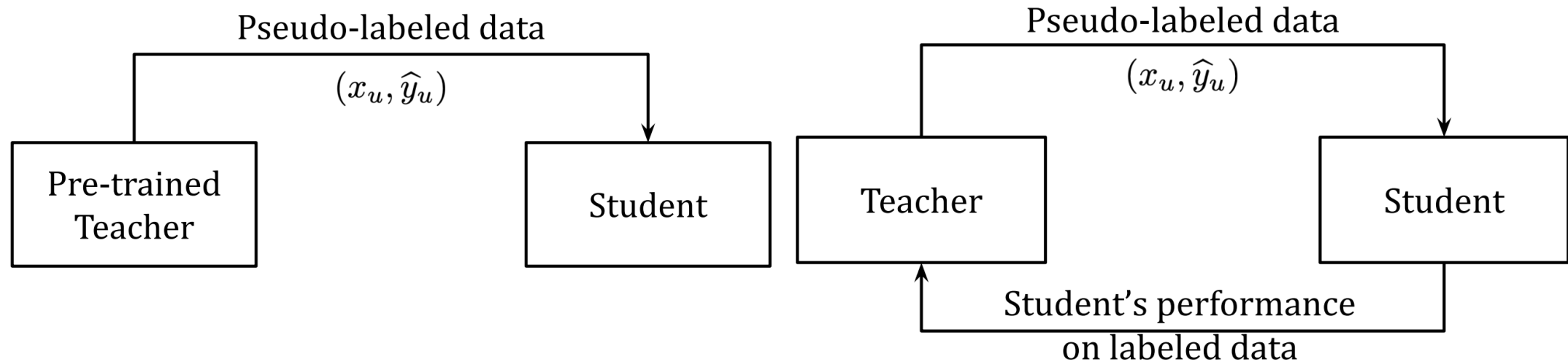


Figure 1: The difference between Pseudo Labels and Meta Pseudo Labels.

Notations

- T, S : teacher network and student network.
 - θ_T, θ_S : their corresponding parameters.
- (x_l, y_l) : a batch of images and their corresponding labels.
- x_u : a batch of unlabeled images.
- $T(x_u; \theta_T)$: the soft predictions of the teacher network on the batch x_u of unlabeled images.
 - $S(x_l; \theta_S), S(x_u; \theta_S)$
- $\text{CE}(q, p)$: cross-entropy loss between q and p .
 - $\text{CE}(y_l, S(x_l; \theta_S))$ is the canonical cross-entropy loss in supervised learning.

Pseudo Labels as an Optimization Problem

- Pseudo Labels (PL) trains the student model to minimize the cross-entropy loss on unlabeled data.

$$\theta_S^{\text{PL}} = \underset{\theta_S}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x_u} \left[\text{CE} \left(T(x_u; \theta_T), S(x_u; \theta_S) \right) \right]}_{:= \mathcal{L}_u(\theta_T, \theta_S)} \quad (1)$$

- θ_S^{PL} always depends on the teacher parameter θ_T via the pseudo targets $T(x_u; \theta_T)$.
- They express the dependency as $\theta_S^{\text{PL}}(\theta_T)$.
- Our purpose: minimize loss on labeled data $\mathcal{L}_l(\theta_S^{\text{PL}}(\theta_T)) := \mathbb{E}_{x_l, y_l} \left[\text{CE} \left(y_l, S(x_l; \theta_S^{\text{PL}}) \right) \right]$.

Pseudo Labels as an Optimization Problem

- Therefore, they could further optimize \mathcal{L}_l w.r.t. θ_T .

$$\begin{aligned} \min_{\theta_T} \quad & \mathcal{L}_l \left(\theta_S^{\text{PL}}(\theta_T) \right), \\ \text{where} \quad & \theta_S^{\text{PL}}(\theta_T) = \operatorname{argmin}_{\theta_S} \mathcal{L}_u(\theta_T, \theta_S). \end{aligned} \tag{2}$$

- They are effectively trying to optimize the teacher on a meta level.
- However, the dependency of $\theta_S^{\text{PL}}(\theta_T)$ on θ_T is extremely complicated.
 - $\nabla_{\theta_T} \theta_S^{\text{PL}}(\theta_T)$ requires unrolling the entire student training process.

Derivation of the Teacher's Update Rule

- See appendix A.
- The result is

$$\nabla_{\theta_T} \mathcal{L}_l = \eta_S \cdot \underbrace{\left(\left(\nabla_{\theta'_S} \text{CE}(y_l, S(x_l; \theta'_S)) \right)^T \cdot \nabla_{\theta_S} \text{CE}(\hat{y}_u, S(x_u; \theta_S)) \right)}_{\text{A scalar } h} \cdot \nabla_{\theta_T} \text{CE}(\hat{y}_u, T(x_u; \theta_T))$$

Unsupervised Data Augmentation

- Teacher's auxiliary losses.

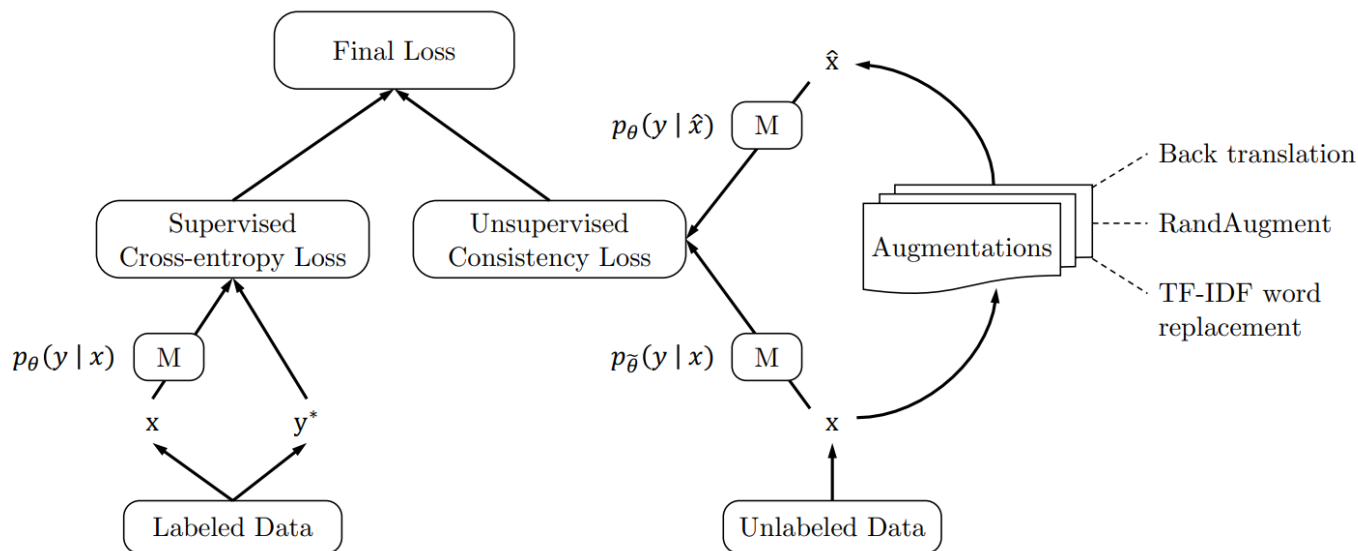


Figure 1: Training objective for UDA, where M is a model that predicts a distribution of y given x .

CIFAR-10 and ImageNet	SVHN
AutoContrast	AutoContrast
Brightness	Brightness
Color	Color
Contrast	Contrast
Equalize	Equalize
Invert	Invert
Sharpness	Sharpness
Posterize	Posterize
Sample Pairing	Solarize
Solarize	ShearX
Rotate	ShearY
ShearX	TranslateY
ShearY	
TranslateX	
TranslateY	

Table 5: Transformations that RandAugment uniformly samples for our datasets.

Pseudo Code for Meta Pseudo Labels with UDA

Algorithm 1 The Meta Pseudo Labels method, applied to a teacher trained with UDA [76].

Input: Labeled data x_l, y_l and unlabeled data x_u .

Initialize $\theta_T^{(0)}$ and $\theta_S^{(0)}$

for $t = 0$ **to** $N - 1$ **do**

 Sample an unlabeled example x_u and a labeled example x_l, y_l

 Sample a pseudo label $\hat{y}_u \sim P(\cdot | x_u; \theta_T)$

 Update the student using the pseudo label \hat{y}_u :

$$\theta_S^{(t+1)} = \theta_S^{(t)} - \eta_S \nabla_{\theta_S} \text{CE}(\hat{y}_u, S(x_u; \theta_S))|_{\theta_S = \theta_S^{(t)}}$$

 Compute the teacher's feedback coefficient as in Equation 12:

$$h = \eta_S \cdot \left(\left(\nabla_{\theta'_S} \text{CE} \left(y_l, S(x_l; \theta_S^{(t+1)}) \right) \right)^\top \cdot \nabla_{\theta_S} \text{CE} \left(\hat{y}_u, S(x_u; \theta_S^{(t)}) \right) \right)$$

 Compute the teacher's gradient from the student's feedback:

$$g_T^{(t)} = h \cdot \nabla_{\theta_T} \text{CE}(\hat{y}_u, T(x_u; \theta_T))|_{\theta_T = \theta_T^{(t)}}$$

 Compute the teacher's gradient on labeled data:

$$g_{T, \text{supervised}}^{(t)} = \nabla_{\theta_T} \text{CE}(y_l, T(x_l; \theta_T))|_{\theta_T = \theta_T^{(t)}}$$

 Compute the teacher's gradient on the UDA loss with unlabeled data:

$$g_{T, \text{UDA}}^{(t)} = \nabla_{\theta_T} \text{CE} \left(\text{StopGradient}(T(\cancel{x_l}); \theta_T), T(\text{RandAugment}(\cancel{x_l}); \theta_T) \right) \Big|_{\theta_T = \theta_T^{(t)}}$$

$\cancel{x_l}$ $\cancel{x_l}$

 Update the teacher:

$$\theta_T^{(t+1)} = \theta_T^{(t)} - \eta_T \cdot \left(g_T^{(t)} + g_{T, \text{supervised}}^{(t)} + g_{T, \text{UDA}}^{(t)} \right)$$

end

return $\theta_S^{(N)}$

▷ Only the student model is returned for predictions and evaluations

Small Scale Experiments

- TwoMoon Experiment.
 - Dataset: [3 (labeled), 997 (unlabeled)] * 2 (classes)
 - Two fc layers (each has 8 units) with sigmoid non-linearity.

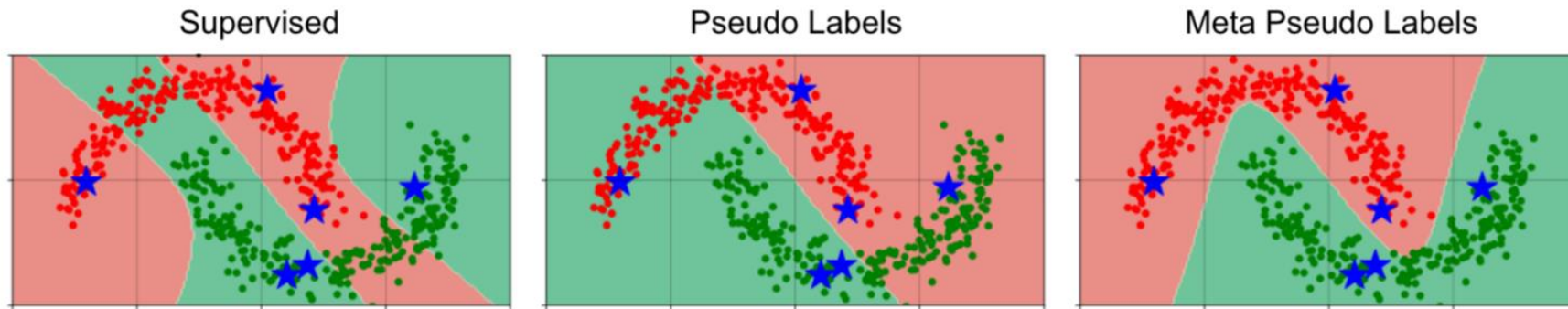


Figure 2: An illustration of the importance of feedback in Meta Pseudo Labels (right). In this example, Meta Pseudo Labels works better than Supervised Learning (left) and Pseudo Labels (middle) on the simple TwoMoon dataset. More details are in Section 3.1.

Small Scale Experiments

- CIFAR-10-4K, SVHN-1K, and ImageNet-10% Experiments.

		CIFAR-10-4K	SVHN-1K	ImageNet-10%	
Method		(mean \pm std)	(mean \pm std)	Top-1	Top-5
Label Propagation Methods	Temporal Ensemble [35]	83.63 \pm 0.63	92.81 \pm 0.27	—	—
	Mean Teacher [64]	84.13 \pm 0.28	94.35 \pm 0.47	—	—
	VAT + EntMin [44]	86.87 \pm 0.39	94.65 \pm 0.19	—	83.39
	LGA + VAT [30]	87.94 \pm 0.19	93.42 \pm 0.36	—	—
	ICT [71]	92.71 \pm 0.02	96.11 \pm 0.04	—	—
	MixMatch [5]	93.76 \pm 0.06	96.73 \pm 0.31	—	—
	ReMixMatch [4]	94.86 \pm 0.04	97.17 \pm 0.30	—	—
	EnAET [72]	94.65	97.08	—	—
	FixMatch [58]	95.74 \pm 0.05	97.72 \pm 0.38	71.5	89.1
	UDA* [76]	94.53 \pm 0.18	97.11 \pm 0.17	68.07	88.19
Self-Supervised Methods	SimCLR [8, 9]	—	—	71.7	90.4
	MOCOv2 [10]	—	—	71.1	—
	PCL [38]	—	—	—	85.6
	PIRL [43]	—	—	—	84.9
	BYOL [21]	—	—	68.8	89.0
Finetuned student model	Meta Pseudo Labels	96.11 \pm 0.07	98.01 \pm 0.07	73.89	91.38
	Supervised Learning with full dataset*	94.92 \pm 0.17	97.41 \pm 0.16	76.89	93.27

Table 2: Image classification accuracy on CIFAR-10-4K, SVHN-1K, and ImageNet-10%.

Small Scale Experiments

- ResNet-50 Experiment

Method	Unlabeled Images	Accuracy (top-1/top-5)
Supervised [24]	None	76.9/93.3
AutoAugment [12]	None	77.6/93.8
DropBlock [18]	None	78.4/94.2
FixRes [68]	None	79.1/94.6
FixRes+CutMix [83]	None	79.8/94.9
NoisyStudent [77]	JFT	78.9/94.3
UDA [76]	JFT	79.0/94.5
Billion-scale SSL [68, 79]	YFCC	82.5/ 96.6
Meta Pseudo Labels	JFT	83.2 /96.5

Table 3: Top-1 and Top-5 accuracy of Meta Pseudo Labels and other representative supervised and semi-supervised methods on ImageNet with ResNet-50.

Large Scale Experiments

- Pushing the Limits of ImageNet Accuracy.

Method	# Params	Extra Data	ImageNet		ImageNet-ReaL [6]
			Top-1	Top-5	Precision@1
ResNet-50 [24]	26M	—	76.0	93.0	82.94
ResNet-152 [24]	60M	—	77.8	93.8	84.79
DenseNet-264 [28]	34M	—	77.9	93.9	—
Inception-v3 [62]	24M	—	78.8	94.4	83.58
Xception [11]	23M	—	79.0	94.5	—
Inception-v4 [61]	48M	—	80.0	95.0	—
Inception-resnet-v2 [61]	56M	—	80.1	95.1	—
ResNeXt-101 [78]	84M	—	80.9	95.6	85.18
PolyNet [87]	92M	—	81.3	95.8	—
SENet [27]	146M	—	82.7	96.2	—
NASNet-A [90]	89M	—	82.7	96.2	82.56
AmoebaNet-A [52]	87M	—	82.8	96.1	—
PNASNet [39]	86M	—	82.9	96.2	—
AmoebaNet-C + AutoAugment [12]	155M	—	83.5	96.5	—
GPipe [29]	557M	—	84.3	97.0	—
EfficientNet-B7 [63]	66M	—	85.0	97.2	—
EfficientNet-B7 + FixRes [70]	66M	—	85.3	97.4	—
EfficientNet-L2 [63]	480M	—	85.5	97.5	—
ResNet-50 Billion-scale SSL [79]	26M	3.5B labeled Instagram	81.2	96.0	—
ResNeXt-101 Billion-scale SSL [79]	193M	3.5B labeled Instagram	84.8	—	—
ResNeXt-101 WSL [42]	829M	3.5B labeled Instagram	85.4	97.6	88.19
FixRes ResNeXt-101 WSL [69]	829M	3.5B labeled Instagram	86.4	98.0	89.73
Big Transfer (BiT-L) [33]	928M	300M labeled JFT	87.5	98.5	90.54
Noisy Student (EfficientNet-L2) [77]	480M	300M unlabeled JFT	88.4	98.7	90.55
Noisy Student + FixRes [70]	480M	300M unlabeled JFT	88.5	98.7	—
Vision Transformer (ViT-H) [14]	632M	300M labeled JFT	88.55	—	90.72
EfficientNet-L2-NoisyStudent + SAM [16]	480M	300M unlabeled JFT	88.6	98.6	—
Meta Pseudo Labels (EfficientNet-B6-Wide)	390M	300M unlabeled JFT	90.0	98.7	91.12
Meta Pseudo Labels (EfficientNet-L2)	480M	300M unlabeled JFT	90.2	98.8	91.02

Table 4: Top-1 and Top-5 accuracy of Meta Pseudo Labels and previous state-of-the-art methods on ImageNet.

Why Should We Use the Student model?

- Generalization: There is room for performance improvement due to labeled data.

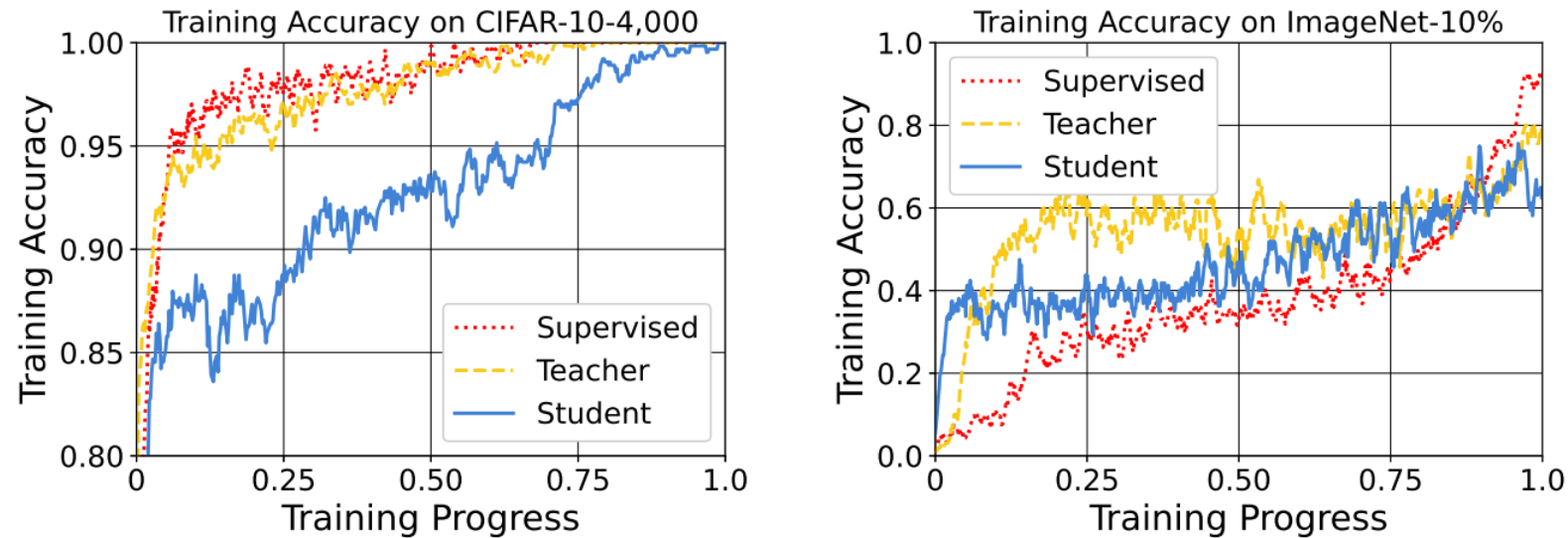


Figure 4: Training accuracy of Meta Pseudo Labels and of supervised learning on CIFAR-10-4,000 and ImageNet-10%. Both the teacher and the student in Meta Pseudo Labels have lower training accuracy, effectively avoiding overfitting.

Thank You