南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Multi-label Iterated Learning for Image Classification with Label Ambiguity

Sai Rajeswar[1,2,3]*, **Pau Rodríguez**[1]*, **Soumye Singhal**[2,3], **David Vazquez**[1], **Aaron Courville**[2,3,4]
[1]Element AI a ServiceNow company, [2]Montréal Institute of Learning Algorithms,
[3]Université de Montréal, [4]CIFAR Fellow
rajsai24@gmail.com, pau.rodriguez@servicenow.com

CVPR 2022

Background

Datasets like ImageNet are weakly labeled since images with multiple object classes present are assigned a single label. This ambiguity biases models towards a single prediction, which could result in the suppression of classes that tend to co-occur in the data.
Softmax cross-entropy results in low multi-label performance since it promotes label exclusiveness. Replacing the softmax with sigmoid activations and casting the output as a set of binary classifiers results in better multi-label validation performance.

ReaL: cleaver, laptop
Sigmoid: laptop
MILe: cleaver, laptop, notebook

ReaL: uniform, riffle
Sigmoid: pickelhaube, riffle
MILe: uniform, pickelhaube, riffle

ReaL: chihuahua, bathtub
Sigmoid: tub
MILe: chihuahua, bathtub, tub

ReaL: polecat, ferret
Sigmoid: ferret
MILe: weasel, polecat, ferret

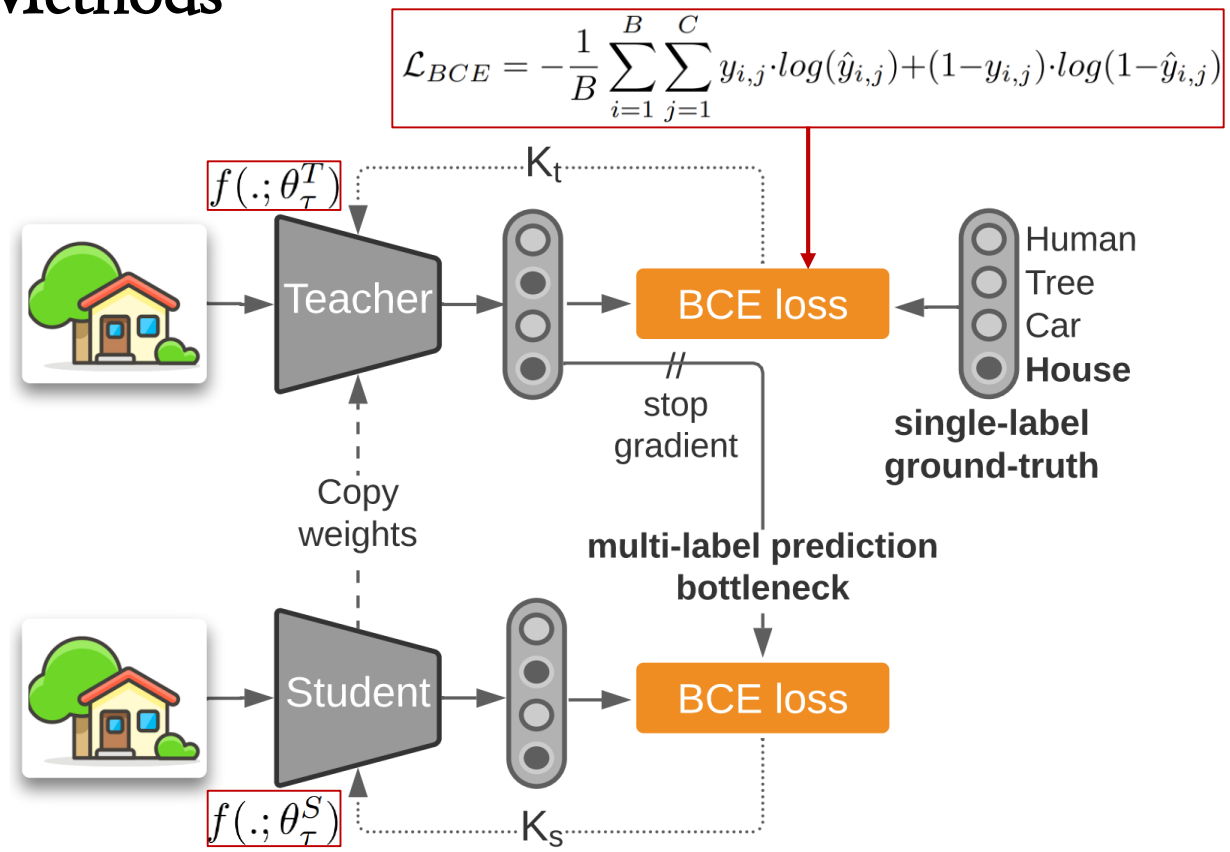
ReaL: schooner, yawl, sandbar, shore
Sigmoid: sandbar
MILe: yawl, sandbar

ReaL: plate, meatloaf
Sigmoid: plate, mashed potato, meat
MILe: plate, mashed potato, meat

ReaL: laptop, notebook
Sigmoid: laptop, desktop computer
MILe: laptop, notebook

ReaL: tabby, keyboard, desk, laptop, monitor, notebook
Sigmoid: keyboard, laptop, notebook
MILe: keyboard, desk, laptop, tabby, notebook

Methods



Algorithm 1 MILe

Require: **Initialize** Student network $\boldsymbol{\theta}_\tau^S$, $\tau = 0$. {*Prepare Iterated Learning*}

repeat
  Copy $\boldsymbol{\theta}_\tau^S$ to $\boldsymbol{\theta}_{\tau+1}^T$ {*Initialize Teacher*}
  for $i = 1$ **to** k_t **do**
    Sample a batch $(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \mathcal{D}_{train}$
    $\hat{\boldsymbol{y}}_i = f_{\boldsymbol{\theta}^T}(\boldsymbol{x_i})$
    $\boldsymbol{\theta}_{\tau+1}^T \leftarrow \boldsymbol{\theta}_{\tau+1}^T + \alpha\nabla\mathcal{L}^{BCE}(\boldsymbol{\theta}_{\tau+1}^T; \boldsymbol{y_i}, \hat{\boldsymbol{y}}_i)$ {*Update* $\boldsymbol{\theta}^T$ *to minimize* L}
  end for {*Finish Interactive Learning*}
  for $i = 1$ **to** k_s **do**
    Sample a batch $(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \mathcal{D}_{train}$
    $\hat{\boldsymbol{y}}_i = \sigma(f_{\boldsymbol{\theta}_{\tau+1}^T}(\boldsymbol{x_i})) > \rho$ {*Generate Pseudo Labels*}
    $\tilde{\boldsymbol{y}}_i = f_{\boldsymbol{\theta}^S}(\boldsymbol{x_i})$
    $\boldsymbol{\theta}_\tau^S \leftarrow \boldsymbol{\theta}_\tau^S + \alpha\nabla\mathcal{L}^{BCE}(\boldsymbol{\theta}_\tau^S; \tilde{\boldsymbol{y}}_i, \hat{\boldsymbol{y}}_i)$ {*Update* $\boldsymbol{\theta}^S$ *to minimize* L}
  end for {*Finish Imitation*}
  Copy $\boldsymbol{\theta}_\tau^S$ to $\boldsymbol{\theta}_{\tau+1}^S$
  $\tau \leftarrow \tau + 1$
until Convergence or maximum τ reached

The interaction phase (teacher learning):

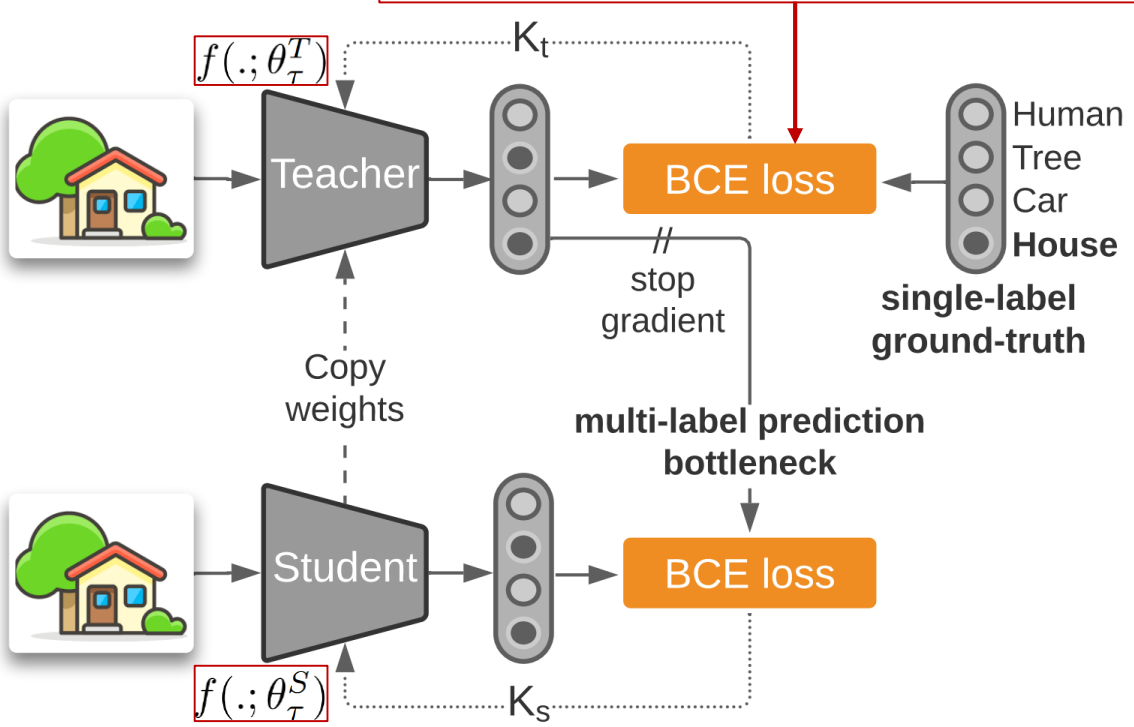
Parameters of the teacher θ_τ^T are initialized using the student parameters θ_τ^S at iteration τ. We train the teacher for k_t learning steps on the labeled images from the dataset, obtaining $f(.;\theta_{\tau+1}^T)$.

The imitation phase (student learning):

Train the student to fit the teacher model for k_S steps, obtaining $f(.;\theta_{\tau+1}^S)$. This is done by training the student on the pseudo labels generated by the teacher on the data. Then, instantiate a new teacher by duplicating the parameters of this new student and iterate the process until convergence.

Methods

$$\mathcal{L}_{BCE} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} y_{i,j}\cdot log(\hat{y}_{i,j}) + (1-y_{i,j})\cdot log(1-\hat{y}_{i,j})$$



Algorithm 1 MILe

Require: **Initialize** Student network $\boldsymbol{\theta}_\tau^S$, $\tau = 0$. {*Prepare Iterated Learning*}

repeat
  Copy $\boldsymbol{\theta}_\tau^S$ to $\boldsymbol{\theta}_{\tau+1}^T$ {*Initialize Teacher*}
  for $i = 1$ **to** k_t **do**
    Sample a batch $(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \mathcal{D}_{train}$
    $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = f_{\boldsymbol{\theta}^T}(\boldsymbol{x_i})$
    $\boldsymbol{\theta}_{\tau+1}^T \leftarrow \boldsymbol{\theta}_{\tau+1}^T + \alpha\nabla\mathcal{L}^{BCE}(\boldsymbol{\theta}_{\tau+1}^T; \boldsymbol{y_i}, \hat{\boldsymbol{y}}_{\boldsymbol{i}})$ {*Update* $\boldsymbol{\theta}^T$ *to minimize* L}
  end for {*Finish Interactive Learning*}
  for $i = 1$ **to** k_s **do**
    Sample a batch $(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \mathcal{D}_{train}$
    $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \sigma(f_{\boldsymbol{\theta}_{\tau+1}^T}(\boldsymbol{x_i})) > \rho$ {*Generate Pseudo Labels*}
    $\tilde{\boldsymbol{y}}_{\boldsymbol{i}} = f_{\boldsymbol{\theta}^S}(\boldsymbol{x_i})$
    $\boldsymbol{\theta}_\tau^S \leftarrow \boldsymbol{\theta}_\tau^S + \alpha\nabla\mathcal{L}^{BCE}(\boldsymbol{\theta}_\tau^S; \tilde{\boldsymbol{y}}_{\boldsymbol{i}}, \hat{\boldsymbol{y}}_{\boldsymbol{i}})$ {*Update* $\boldsymbol{\theta}^S$ *to minimize* L}
  end for {*Finish Imitation*}
  Copy $\boldsymbol{\theta}_\tau^S$ to $\boldsymbol{\theta}_{\tau+1}^S$
  $\tau \leftarrow \tau + 1$
until Convergence or maximum τ reached

In order to prevent the student from overfitting the teacher, we restrict the amount of training updates for each of the modules.
Formally, let N be the size of the dataset, k_t be the number of training iterations of the teacher, and k_s the number of student iterations. In general, we set $k_t \ll N$ to prevent the teacher from overfitting one-hot labels and $k_s \leq k_t$ to prevent the student from overfitting the teacher.

Experiments

Train a ResNet-18 and a ResNet-50 model.
compare three different methods.
(i) Softmax: standard softmax cross-entropy loss used to train the original ResNet backbone.
(ii) Sigmoid: we substitute the cross-entropy loss for a binary cross-entropy (BCE) loss.
(iii) MILe: the proposed method
Label coverage: indicates the total fraction of labels per sample predicted by the multi-label classifier.

	ImageNet fraction:	1%	5%	10%	100%	1%	5%	10%	100%
Metric	Method	ResNet-50				ResNet-18			
Accuracy	Softmax	6.32	36.71	53.50	76.33	6.61	31.5	48.82	70.41
	Sigmoid	6.70	36.9	55.01	76.35	6.88	31.1	49.14	70.46
	MILe (ours)	**9.10**	**42.52**	**57.29**	**77.12**	**8.2**	**36.2**	**51.31**	**71.12**
ReaL-Acc	Softmax	7.19	42.55	60.21	82.76	8.80	35.88	55.11	77.77
	Sigmoid	8.38	46.04	62.96	83.22	9.04	37.66	57.52	81.01
	MILe (ours)	**11.5**	**48.36**	**65.42**	**83.75**	**9.18**	**41.65**	**58.57**	**81.52**
ReaL-F1	Softmax	6.77	40.51	57.33	78.5	8.28	34.20	52.51	73.83
	Sigmoid	7.17	41.11	58.46	78.61	8.39	33.56	52.12	73.85
	MILe (ours)	**10.76**	**45.02**	**62.11**	**79.89**	**8.55**	**38.49**	**53.8**	**74.48**
Label Coverage	Softmax	1.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Sigmoid	1.09	1.11	1.10	1.11	1.07	1.10	1.15	1.15
	MILe (ours)	1.05	1.08	1.09	1.16	1.06	1.07	1.12	1.17

Experiments

These results suggest that the iterated learning bottleneck acts as a regularizer that prevents the model from learning noisy labels which are more difficult to fit. Noise memorization happens later in the training procedure.

Method	Architecture	WebVision		ImageNet	
		Top-1	Top-5	Top-1	Top-5
CrossEntropy [63]	ResNet-50	66.4	83.4	57.7	78.4
MentorNet [30]	InceptionRes-V2	70.8	88.0	62.5	83.0
CurriculumNet [23]	Inception-V2	72.1	89.1	64.8	84.9
CleanNet [37]	ResNet-50	70.3	87.8	63.4	84.6
CurriculumNet [23, 63]	ResNet-50	70.7	88.6	62.7	83.4
SOM [63]	ResNet-50	72.2	89.5	65.0	85.1
Distill [72]	ResNet-50	-	-	65.8	**85.8**
MoPro (dec.) [39]	ResNet-50	72.4	89.0	65.7	85.1
Multimodal [56]	Inception-V3	73.15	89.73	-	-
Sigmoid	ResNet-50	72.1	89.5	65.4	85.0
MILe (ours)	ResNet-50	**75.2**	**90.3**	**67.1**	85.6
Initial Vanilla Model	ResNet-50-D	75.08	89.22	67.23	84.09
SCC [67]	ResNet-50-D	75.36	89.38	67.93	84.77
SCC+GBA [67]	ResNet-50-D	75.69	89.42	68.35	85.24
MILe (ours)	ResNet-50-D	**76.5**	**90.9**	**68.7**	**86.4**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Self-supervised Fine-tuning:

We explore whether iterated learning improves the performance of self-supervised models in the fully- and semi-supervised fine-tuning regimes.

Method	ImageNet Validation			ImageNet ReaL-F1		
	1%	10%	100%	1%	10%	100%
SimCLR [13]	48.3	65.6	76.25	51.54	69.16	76.91
BYOL [21]	53.2	68.8	77.2	54.32	70.81	78.85
SwAV [11]	53.9	70.2	77.74	55.79	71.22	79.18
MoCo-v2 [15]	51.72	66.5	77.12	53.34	70.75	79.04
MILe (Ours) + [15]	**52.62**	**67.4**	**77.38**	**56.08**	**71.48**	**80.03**
SimCLR-v2-sk0 [14]	58.18	68.9	76.3	57.25	70.11	78.83
MILe (Ours) + [14] (sk0)	**61.85**	**70.5**	**77.29**	**60.49**	**72.76**	**79.38**
SimCLR-v2-sk1 [14]	64.7	72.4	78.7	62.77	74.21	79.43
MILe (Ours) + [14] (sk1)	**69.4**	**74.7**	**79.5**	**65.04**	**76.40**	**81.53**

Semi-supervised learning:

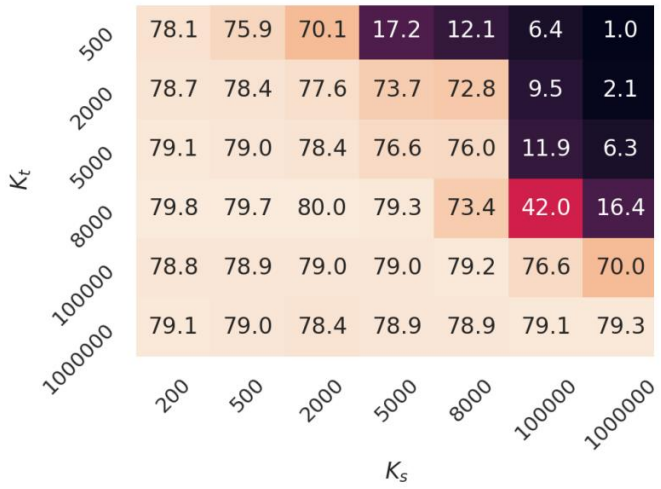
ImageNet top-1 accuracy for ResNet-50 (R50) distilled from a SimCLR model.
2×: teacher has 2× parameters than the student.

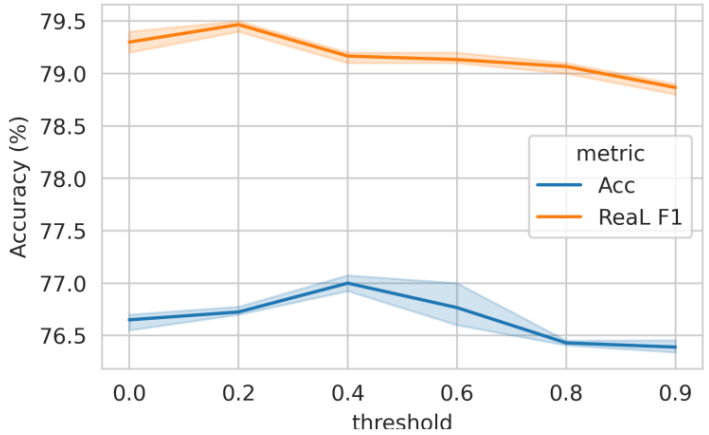
Method	Teacher	Label fraction	
		1%	10%
Distilled [14]	R50 (2×+SK)	69.0	75.1
Self-distilled [14]	R50 (1x+SK)	70.15	74.43
MILe (ours)	R50 (1x+SK)	**73.08**	**75.3**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Ablation Study :





This is expected since a small k_t would let the imitation phase constantly disrupt supervised learning via interaction with the data, while a large k_t does not reap the benefits of distillation. For a given k_t we find that the optimal k_s lies in the mid-range and the other way around.

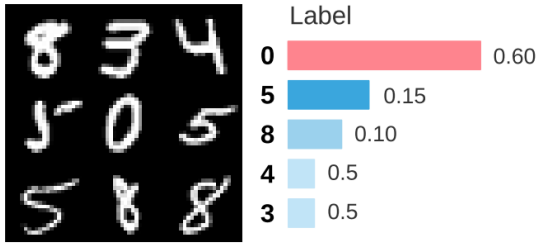
Lower thresholds bias the student towards producing multi-label outputs, even for low-confidence classes. Larger threshold values make the student tend towards single-labeled prediction, only predicting labels for which the confidence is high.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Explore the limits of MILe:

The center digit has a probability of 0.6 to be chosen as the label for the whole grid.
For the results on multi-label MNIST. The first column displays the F1 score when the threshold for positive labels is set to 0.25 and the second column shows the F1 score for a threshold of 0.5.



	F1@0.25	F1@0.5
Softmax	28.69	28.69
Sigmoid	29.10	28.67
MILe (ours)	**41.35**	**34.32**

ReaL label recovery:

Secondary label recovery. Mean average precision over labels that appear in ReaL but not in the original ImageNet validation set.

Method	ResNet-50		ResNet-18	
	10% data	100% data	10% data	100% data
Softmax	0.2171	0.2679	0.1983	0.2648
Sigmoid	0.2310	0.2845	0.2047	0.2836
MILe (ours)	**0.3042**	**0.3248**	**0.2187**	**0.2880**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Noise multi-label Dataset:

Comparison on CelebA multi-attribute classification. Just as in ReaL ImageNet validation, we use F1-score (based on the intersection over union) measure to evaluate the methods.

Method	F1-score
CE-Sigmoid	80.14
ResNet-18(FPR) [8]	77.55
ResNet-34 (FPR) [8]	79.96
MILe (ours)	**81.40**

	5 o Clock Shadow	*Arched Eyebrows*	*Attractive*	*Bags Under Eyes*	*Bald*	*Bangs*	*Big Lips*	*Big Nose*	*Black Hair*	*Blond Hair*	*Blurry*	*Brown Hair*	*Bushy Eyebrows*	*Chubby*	*Double Chin*	*Eyeglasses*	*Goatee*	*Gray Hair*	*Heavy Makeup*	*High Cheekbones*
Triplet-kNN [55]	66	73	83	63	75	81	55	68	82	81	43	76	68	64	60	82	73	72	88	86
PANDA [71]	76	77	85	67	74	92	56	72	84	91	50	**85**	74	65	64	88	84	79	95	**89**
Anet [44]	81	76	**87**	70	73	90	57	**78**	**90**	90	56	83	82	70	68	95	**86**	85	**96**	**89**
MILe	**85**	**83**	82	**74**	**82**	**92**	**65**	74	88	**91**	**76**	79	**83**	**72**	**72**	**98**	**86**	**86**	93	**89**

	Male	*Mouth Slightly Open*	*Mustache*	*Narrow Eyes*	*No Beard*	*Oval Face*	*Pale Skin*	*Pointy Nose*	*Receding Hairline*	*Rosy Cheeks*	*Sideburns*	*Smiling*	*Straight Hair*	*Wavy Hair*	*Wearing Earrings*	*Wearing Hat*	*Wearing Lipstick*	*Wearing Necklace*	*Wearing Necktie*	*Young*
Triplet-kNN [55]	91	92	57	47	82	61	63	61	60	64	71	92	63	77	69	84	91	50	73	75
PANDA [71]	**99**	93	63	51	87	66	69	67	67	68	81	**98**	66	78	77	90	**97**	51	**85**	78
Anet [44]	**99**	**96**	61	57	93	**67**	**77**	69	70	**76**	79	97	69	81	83	90	95	**59**	79	**84**
MILe	**99**	95	**74**	**77**	**94**	64	75	**69**	**77**	74	**87**	94	**74**	**83**	**84**	**94**	93	56	77	81

Table 9. Mean per-class balanced accuracy in percentage points for each of the 40 face attributes on CelebA.

Thanks