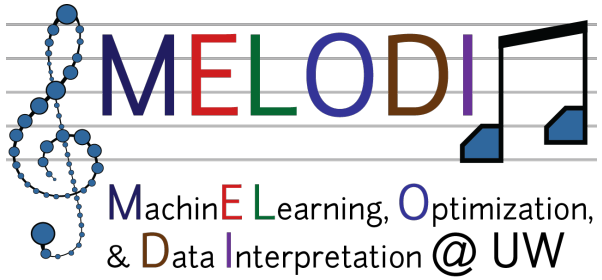
Curriculum Learning by Dynamic Instance Hardness

Tianyi Zhou*, Shengjie Wang*, Jeff A. Bilmes

University *of* Washington, Seattle



Curriculum Learning with Dynamic Instance Hardness (DIH)

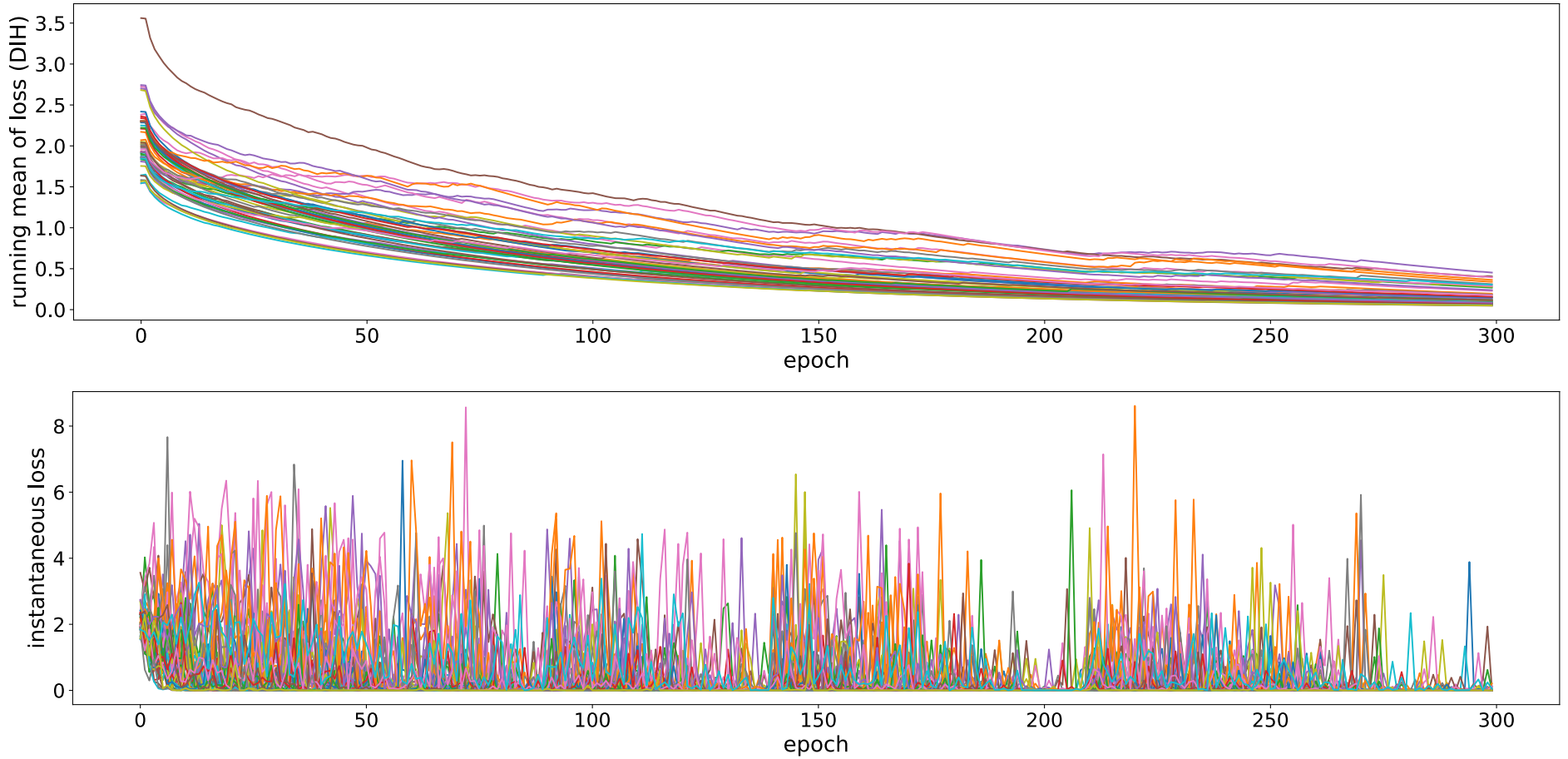
- Curriculum Learning (CL): Instead of randomly selecting training samples per mini-batch, select samples based on some criteria (a curriculum) to better guide the training process.
- Previous Curriculum Learning Strategies:
  - Prefer easy samples, criteria: losses of samples at the current training step.

- Dynamic Instance Hardness
  - Use a running mean of the instantaneous hardness metric (e.g., loss) as the criterion of data selection:

$$r_{t+1}(i) = \begin{cases} \gamma \times a_t(i) + (1-\gamma) \times r_t(i) & \text{if } i \in S_t \\ r_t(i) & \text{else ,} \end{cases}$$

DIH metric v.s. Previous CL metric

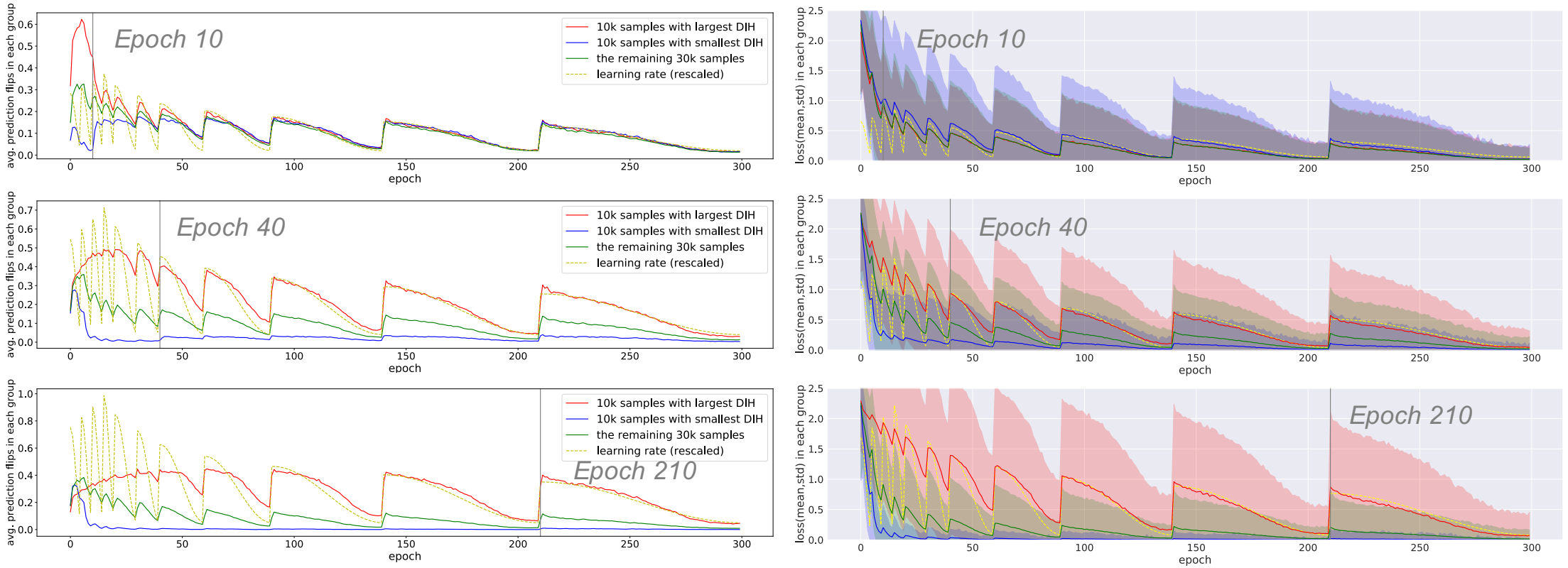
DIH is more smooth/consistent over training epochs than instantaneous hardness.



Top: DIH vs. Bottom: instantaneous loss of 50 randomly selected samples from CIFAR10 on WideResNet-28-10.

DIH to identify memorable/forgettable samples

DNNs have very different training dynamics on samples with small and large DIH.



LEFT: Averaged prediction-flip and RIGHT: losses of the three groups of samples partitioned by a DIH metric computed at epoch 10,40 and 210. DIH in early stage (Epoch 40) can predict the forgettable/memorable samples for later stages.

DIH Curriculum Learning (DIHCL)

- At every training step, prefer samples with large DIH metric values (prefer hard samples and reduce computation on already learnt samples).
- Has theoretical guarantee if we make assumptions about the class of function that generates the DIH metrics.

Algorithm 1 DIH Curriculum Learning (DIHCL-Greedy)

input: $\{(x_i, y_i)\}_{i=1}^n$, $\pi(\cdot;\eta)$, $\ell(\cdot,\cdot)$, $F(\cdot;w)$;
$\eta_{1:T}$; T, T_0; $\gamma, \gamma_k \in [0,1]$
initialize: w, η_1, $k_1 = n$, $r_0(i) = 1\ \forall i \in [n]$
for $t \in \{1, \cdots, T\}$ **do**
  if $t \leq T_0$ **then**
    $S_t \leftarrow [n]$;
  else
    Let $S_t = \text{argmax}_{S:|S|=k_t} \sum_{i \in S} r_t(i)$;
  end if
  Apply optimization $\pi(\cdot;\eta)$ to update model:

$$w_t \leftarrow w_{t-1} + \pi\left(\nabla_w \sum_{i \in S_t} \ell(y_i, F(x_i; w_{t-1})); \eta_t\right)$$

  Compute normalized $a_t(i)$ for $i \in S_t$ using Eq. (2);
  Update DIH $r_{t+1}(i)$ using Eq. (1);
  $k_{t+1} \leftarrow \gamma_k \times k_t$;
end for

Experiment Results

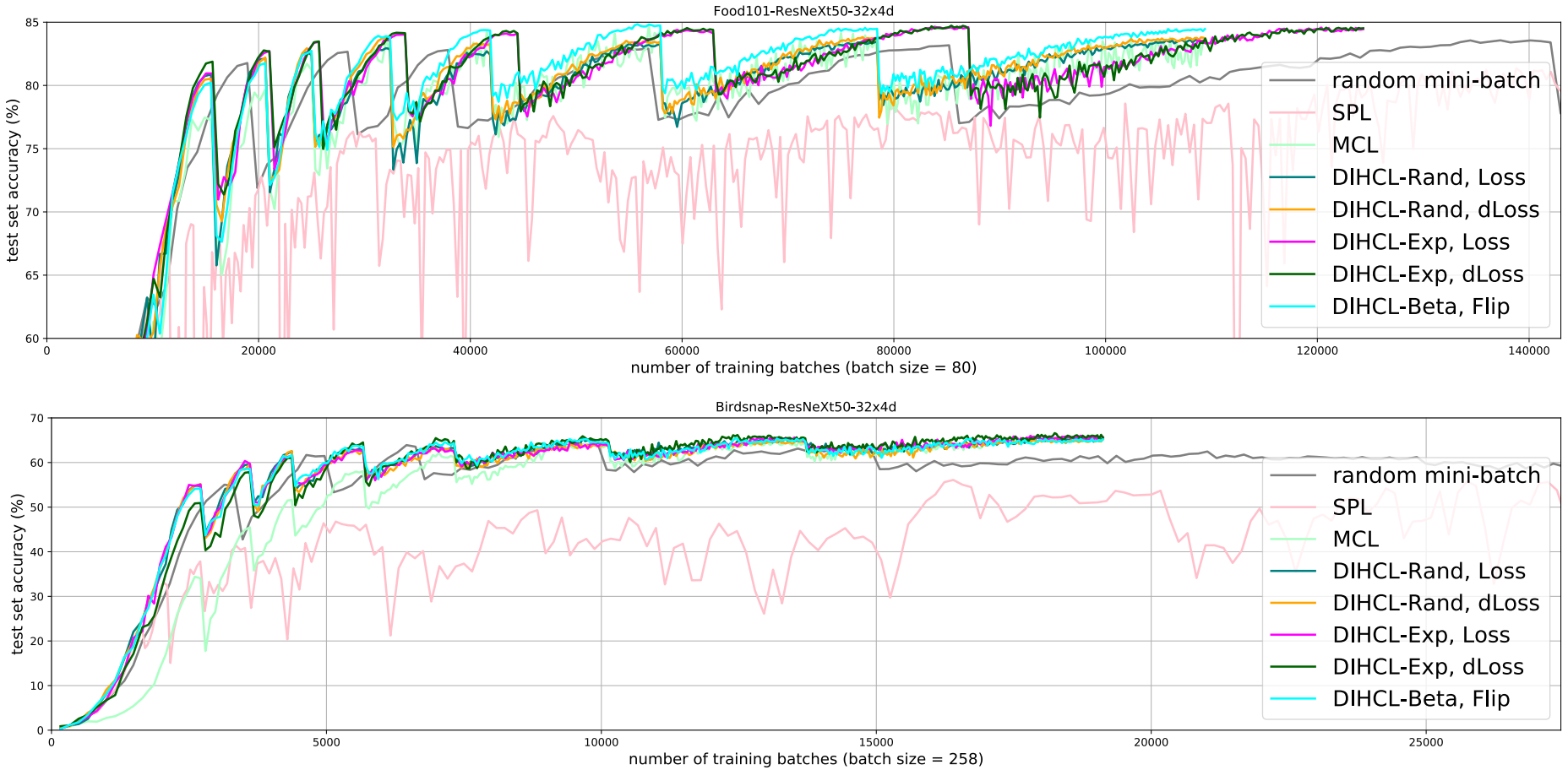
- DIHCL outperforms baseline CL methods on multiple datasets.

Curriculum	CIFAR10	CIFAR100	Food-101	ImageNet	STL10	SVHN	KMNIST	FMNIST	Birdsnap	Aircraft	Cars
Rand mini-batch	96.18	79.64	83.56	75.04	86.06	96.48	98.67	95.22	64.23	74.71	78.73
SPL	93.55	80.25	81.36	73.23	81.33	96.15	97.24	92.09	63.26	68.95	77.61
MCL	96.60	80.99	84.18	75.09	88.57	96.93	99.09	95.07	65.76	75.28	76.98
DIHCL-Rand, Loss	96.76	80.77	83.82	75.41	87.25	96.81	99.10	95.69	65.62	79.00	80.91
DIHCL-Rand, dLoss	96.73	80.65	83.82	75.34	86.93	96.83	99.14	95.64	65.25	79.93	78.70
DIHCL-Exp, Loss	97.03	82.23	84.65	75.10	88.36	96.91	99.20	95.45	66.13	77.68	79.85
DIHCL-Exp, dLoss	96.40	81.42	84.75	75.62	89.41	96.80	99.18	95.50	66.59	79.72	81.48
DIHCL-Beta, Flip	96.51	81.06	84.94	76.33	86.88	97.18	99.05	95.66	65.48	78.49	80.13

For each dataset, the best accuracy is in blue, the second best is red, and third best green.

Experiment Results

DIHCL reaches the best performance faster.



The top plot is on dataset Food-101, and the bottom plot is on dataset Birdsnap.

Thank you!