A Unified Approach to Domain Incremental Learning with Memory: Theory and Algorithm

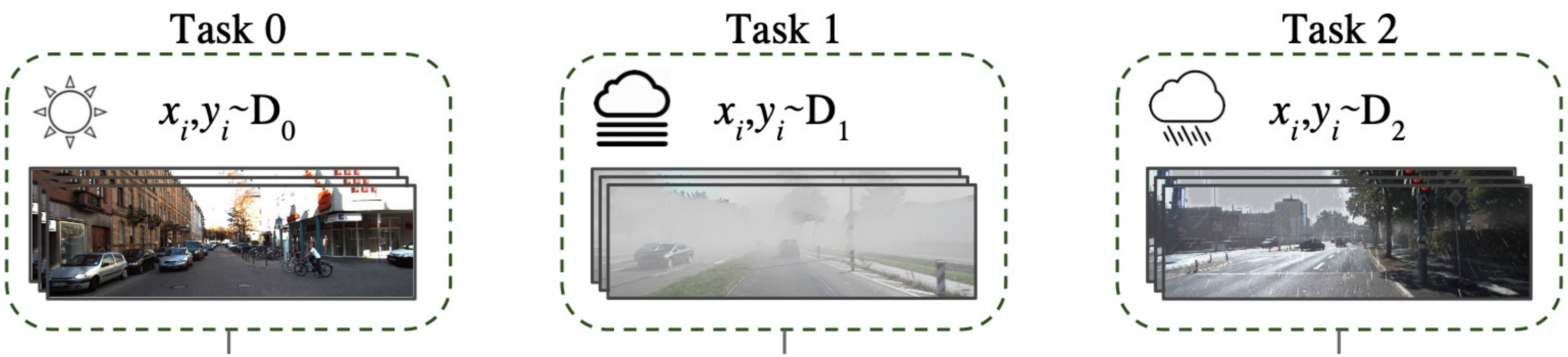
Haizhou Shi, Hao Wang

Computer Science Department, Rutgers University

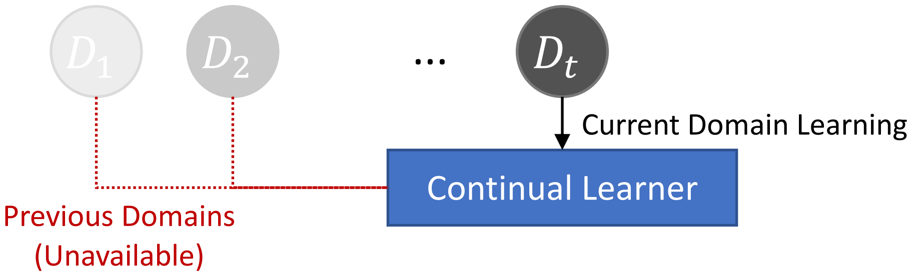
Background

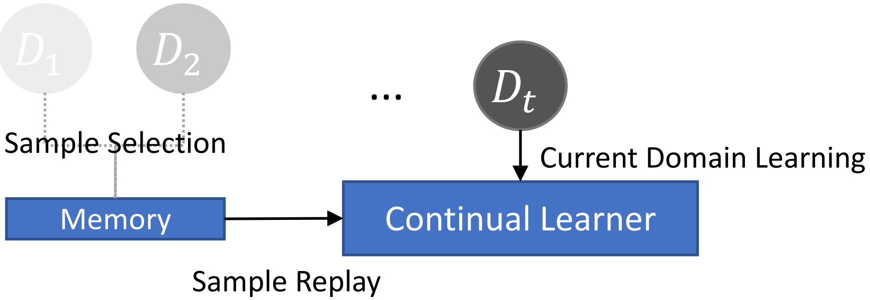
RUTGERS UNIVERSITY

- Domain Incremental Learning (DIL)
  - Machine learning models are required to incrementally learn the evolving data distributions.
  - E.g., autonomous driving under different weather conditions.



  - Memory constraint: no (or very limited size of) the past data can be stored during training.

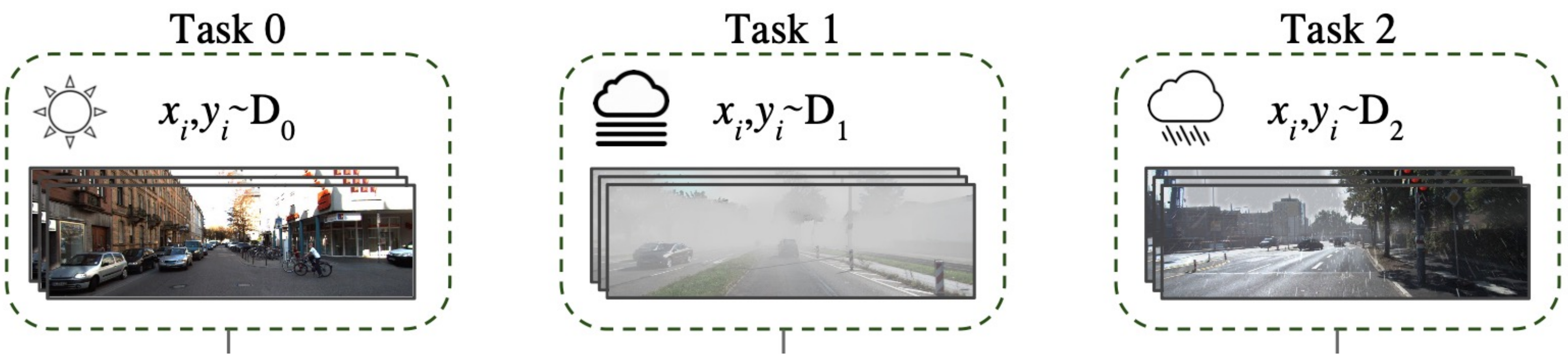




Background

RUTGERS UNIVERSITY

- Domain Incremental Learning (DIL)
  - Machine learning models are required to incrementally learn the evolving data distributions.
  - E.g., autonomous driving under different weather conditions.

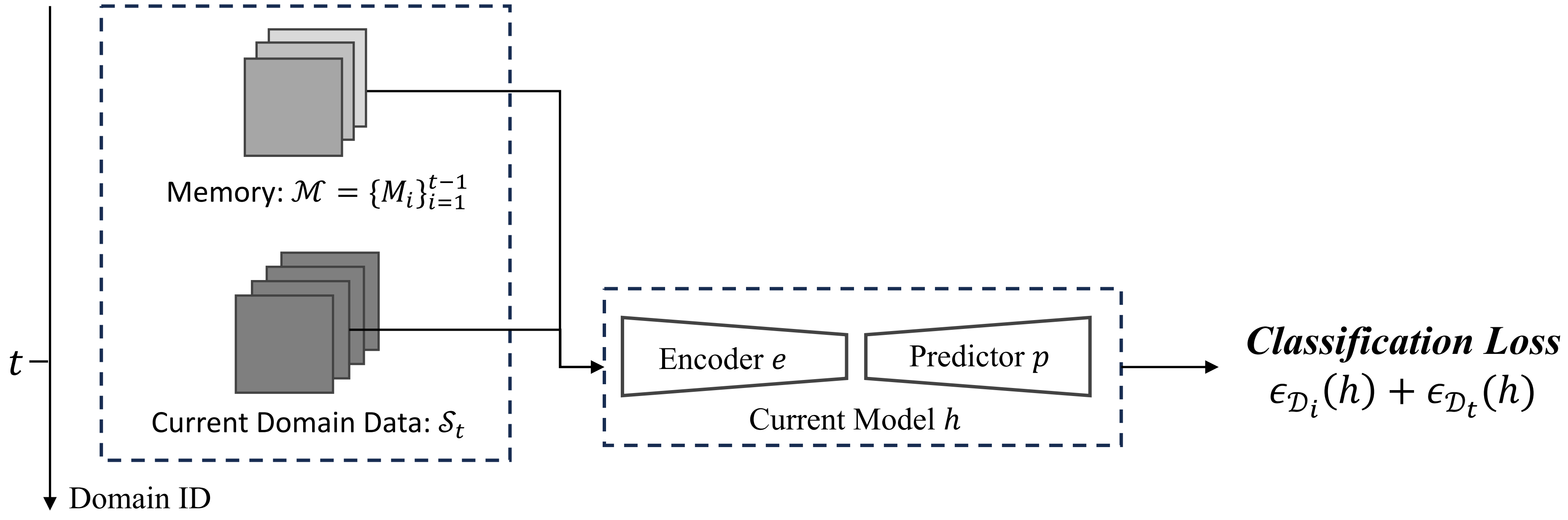


  - Memory constraint: no (or very limited size of) the past data can be stored during training.
  - Goal of DIL: minimize the model's risk over ***all domains***.

$$\mathcal{L}^*(\theta) = \mathcal{L}_t(\theta) + \mathcal{L}_{1:t-1}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}[\ell(y, h_\theta(x)] + \sum_{i=1}^{t-1} \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell(y, h_\theta(x)]$$

ERM-Based Generalization Bound

- Empirical Risk Minimization (ERM) via Experience Replay (ER)

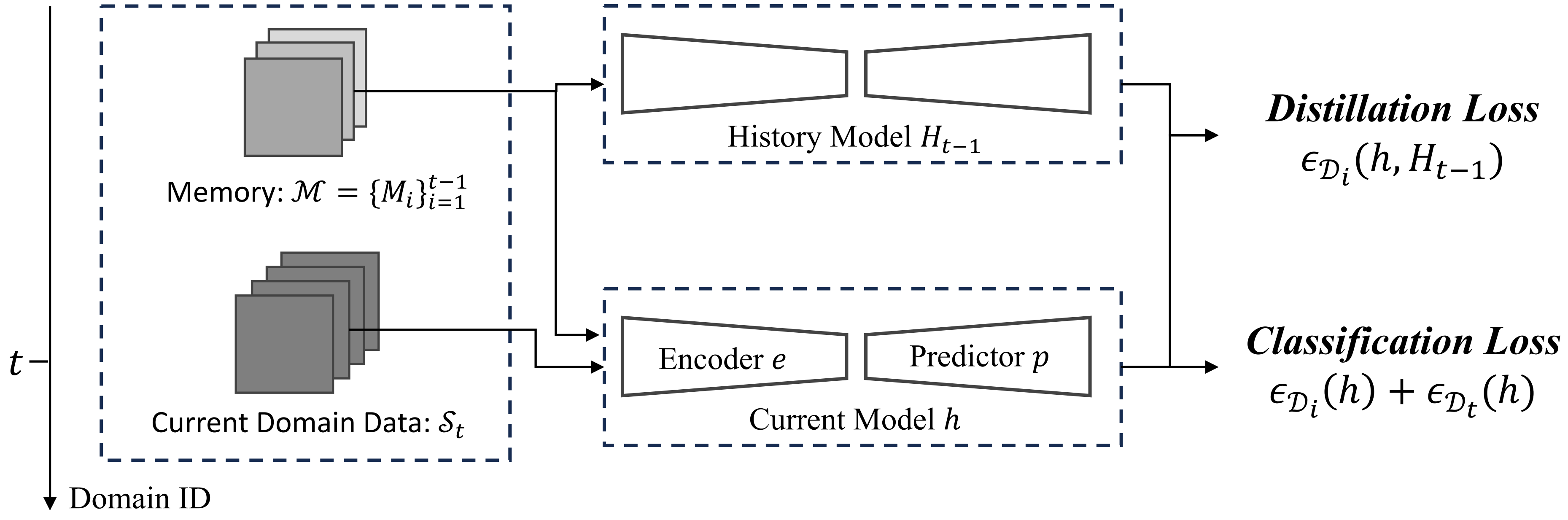


- [Lemma 3.1] Trivially replaying the memory will cause ***a loose generalization bound***.

$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \leq \sum_{i=1}^{t} \widehat{\epsilon}_{\mathcal{D}_i}(h) + \sqrt{\left(\frac{1}{N_t} + \sum_{i=1}^{t-1} \frac{1}{\widetilde{N}_i}\right)\left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}.$$

Intra-Domain Model-Based Bound

- Dark Experience Replay (DER++)



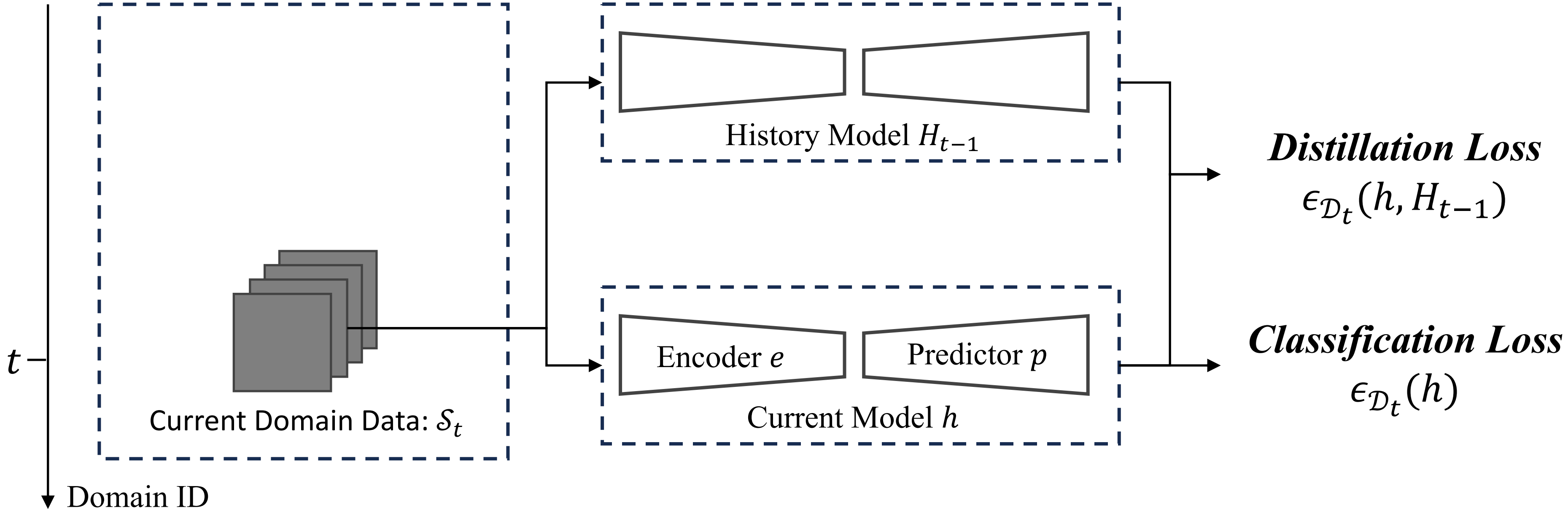
- [Lemma 3.2] Intra-Domain Model-Based Bound

$$\epsilon_{\mathcal{D}_i}(h) \leq \epsilon_{\mathcal{D}_i}(h, H_{t-1}) + \epsilon_{\mathcal{D}_i}(H_{t-1}),$$

Cross-Domain Model-Based Bound

RUTGERS UNIVERSITY

- Learning without Forgetting (LwF)



- [Lemma 3.3] Cross-Domain Model-Based Bound

$$\epsilon_{\mathcal{D}_i}(h) \leq \epsilon_{\mathcal{D}_t}(h, H_{t-1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \epsilon_{\mathcal{D}_i}(H_{t-1}),$$

UDIL: A *Unified* Bound for DIL

- A set of coefficients $\{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{t-1}$ (with $\alpha_i + \beta_i + \gamma_i = 1$) integrates them into one unified bound.

$\epsilon_{\mathcal{D}_i}(h) \rightarrow$

- $\gamma_i \epsilon_{\mathcal{D}_i}(h)$ — ***Naïve ERM***
- $\alpha_i[\epsilon_{\mathcal{D}_i}(h, H_{t-1}) + \epsilon_{\mathcal{D}_i}(H_{t-1})]$ — ***Intra-Domain Bound***
- $\beta_i[\epsilon_{\mathcal{D}_i}(h, H_{t-1}) + \epsilon_{\mathcal{D}_t}(h, H_{t-1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t)]$ — ***Cross-Domain Bound***

- [Theorem 3.4] Unified Generalization Bound for all domains

$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \widehat{\epsilon}_{\mathcal{D}_i}(h) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_i}(h, H_{t-1})] \right\} + \left\{ \widehat{\epsilon}_{\mathcal{D}_t}(h) + (\sum_{i=1}^{t-1} \beta_i) \widehat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\}$$
$$+ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(H_{t-1})$$
$$+ \sqrt{\left(\frac{(1+\sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\widetilde{N}_i} \right) \left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}$$

UDIL: A *Unified* Bound for DIL

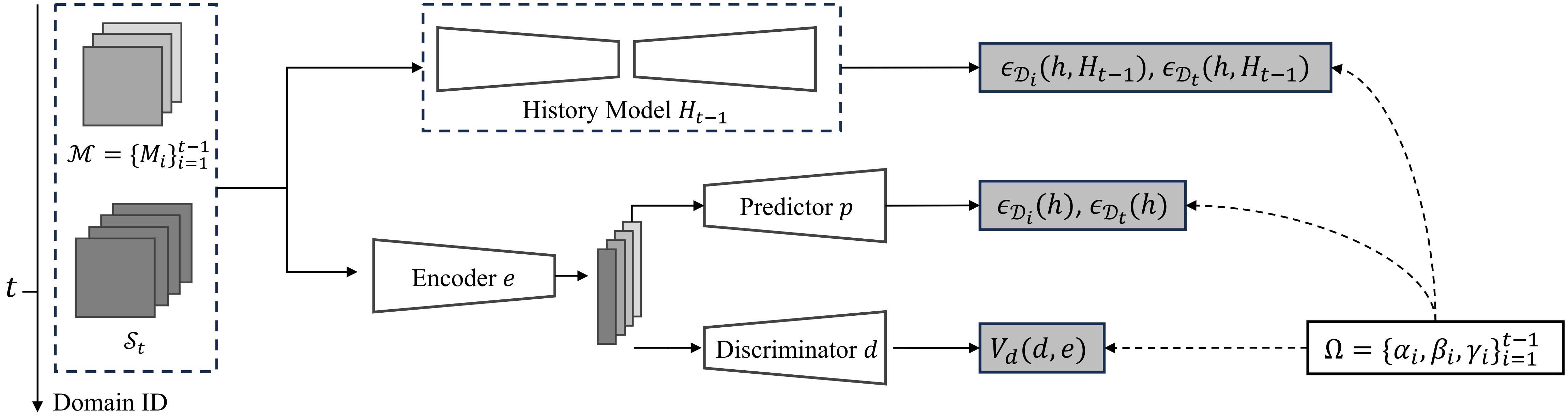
- UDIL ***unifies*** multiple existing methods under certain conditions.

	α_i	β_i	γ_i	Transformed Objective	Condition
UDIL (Ours)	$[0,1]$	$[0,1]$	$[0,1]$	-	-
LwF [52]	0	1	0	$\mathcal{L}_{\text{LwF}}(h) = \widehat{\ell}_{\mathcal{X}_t}(h) + \lambda_o \widehat{\ell}_{\mathcal{X}_t}(h, H_{t-1})$	$\lambda_o = t-1$
ER [75]	0	0	1	$\mathcal{L}_{\text{ER}}(h) = \widehat{\ell}_{B_t}(h) + \sum_{i=1}^{t-1} \frac{\lvert B'_t\rvert/(t-1)}{\lvert B_t\rvert} \widehat{\ell}_{B'_i}(h)$	$\lvert B_t\rvert = \frac{\lvert B'_t\rvert}{(t-1)}$
DER++ [8]	1/2	0	1/2	$\mathcal{L}_{\text{DER++}}(h) = \widehat{\ell}_{B_t}(h) + \frac{1}{2}\sum_{i=1}^{t-1} \frac{\lvert B'_t\rvert/(t-1)}{\lvert B_t\rvert} [\widehat{\ell}_{B'_i}(h) + \widehat{\ell}_{B'_i}(h, H_{t-1})]$	$\lvert B_t\rvert = \frac{\lvert B'_t\rvert}{(t-1)}$
iCaRL [74]	1	0	0	$\mathcal{L}_{\text{iCaRL}}(h) = \widehat{\ell'}_{B_t}(h) + \sum_{i=1}^{t-1} \frac{\lvert B'_t\rvert/(t-1)}{\lvert B_t\rvert} \widehat{\ell'}_{B'_i}(h, H_{t-1})$	$\lvert B_t\rvert = \frac{\lvert B'_t\rvert}{(t-1)}$
CLS-ER [4]	$\frac{\lambda}{\lambda+1}$	0	$\frac{1}{\lambda+1}$	$\mathcal{L}_{\text{CLS-ER}}(h) = \widehat{\ell}_{B_t}(h) + \sum_{i=1}^{t-1} \frac{1}{t-1} \widehat{\ell}_{B'_i}(h) + \sum_{i=1}^{t-1} \frac{\lambda}{t-1} \widehat{\ell}_{B'_i}(h, H_{t-1})$	$\lambda = t-2$
ESM-ER [80]	$\frac{\lambda}{\lambda+1}$	0	$\frac{1}{\lambda+1}$	$\mathcal{L}_{\text{ESM-ER}}(h) = \widehat{\ell}_{B_t}(h) + \sum_{i=1}^{t-1} \frac{1}{r(t-1)} \widehat{\ell}_{B'_i}(h) + \sum_{i=1}^{t-1} \frac{\lambda}{r(t-1)} \widehat{\ell}_{B'_i}(h, H_{t-1})$	$\begin{cases} \lambda = -1 + r(t-1) \\ r = 1 - e^{-1} \end{cases}$
BiC [100]	$\frac{t-1}{2t-1}$	$\frac{t-1}{2t-1}$	$\frac{1}{2t-1}$	$\mathcal{L}_{\text{BiC}}(h) = \widehat{\ell}_{B_t}(h) + \sum_{i=1}^{t-1} \frac{(t-1)\lvert B_i\rvert}{\lvert B_t\rvert} \widehat{\ell}_{B'_i}(h, H_{t-1})$ $+ (t-1)\widehat{\ell}_{B_t}(h, H_{t-1}) + \sum_{i=1}^{t-1} \frac{\lvert B_i\rvert}{\lvert B_t\rvert} \widehat{\ell}_{B'_i}(h)$	$\lvert B_i\rvert = \lvert B_t\rvert$

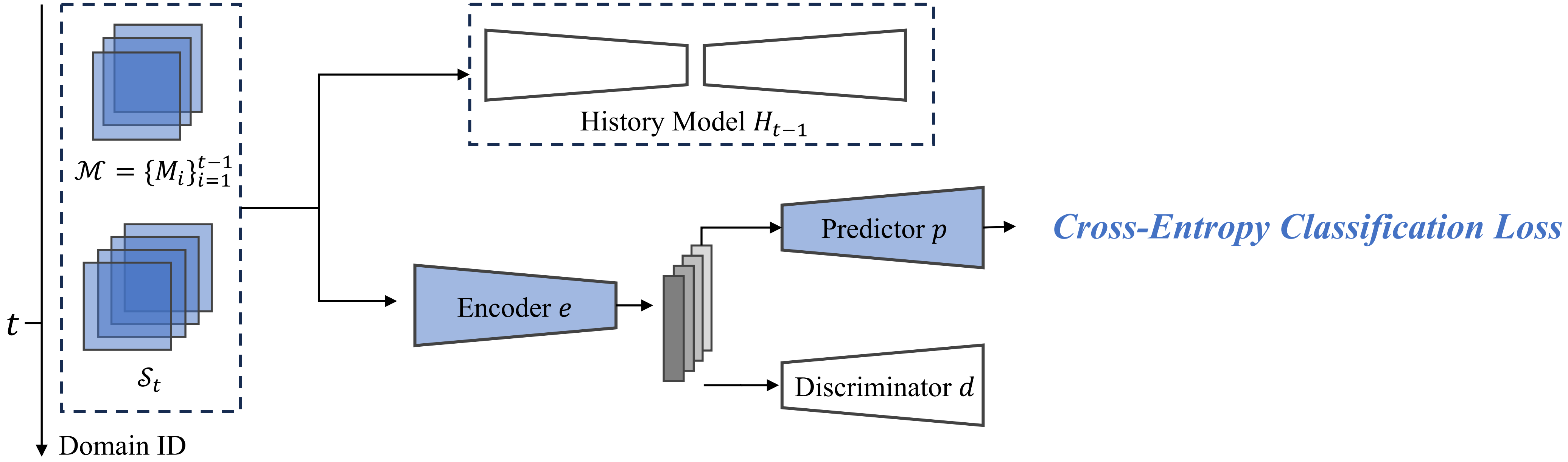
UDIL: An *Adaptive* Bound for DIL

RUTGERS UNIVERSITY

- UDIL can ***adaptively*** adjust the coefficients based on the data and the history model H_{t-1}.
- It will, ideally, minimize the ***tightest bound*** in the family of all the generalization bounds.



UDIL: An *Adaptive* Bound for DIL



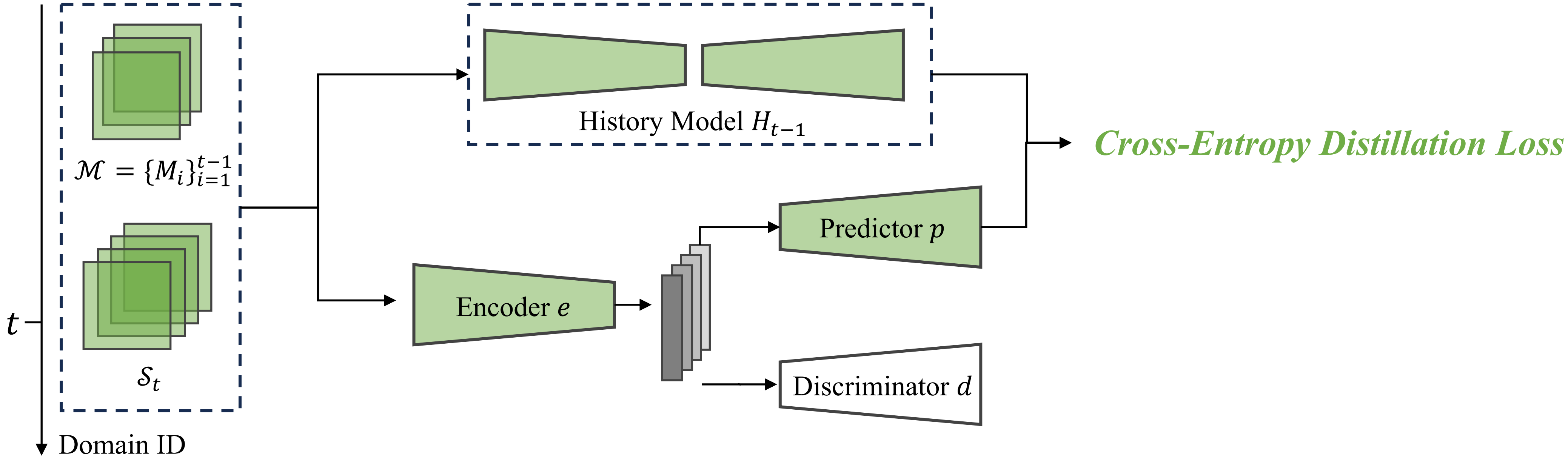
$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \widehat{\epsilon}_{\mathcal{D}_i}(h) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_i}(h, H_{t-1})] \right\} + \left\{ \widehat{\epsilon}_{\mathcal{D}_t}(h) + (\sum_{i=1}^{t-1} \beta_i) \widehat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\}$$

$$+ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(H_{t-1})$$

$$+ \sqrt{\left(\frac{(1+\sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\widetilde{N}_i} \right) \left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}$$

UDIL: An *Adaptive* Bound for DIL

RUTGERS UNIVERSITY

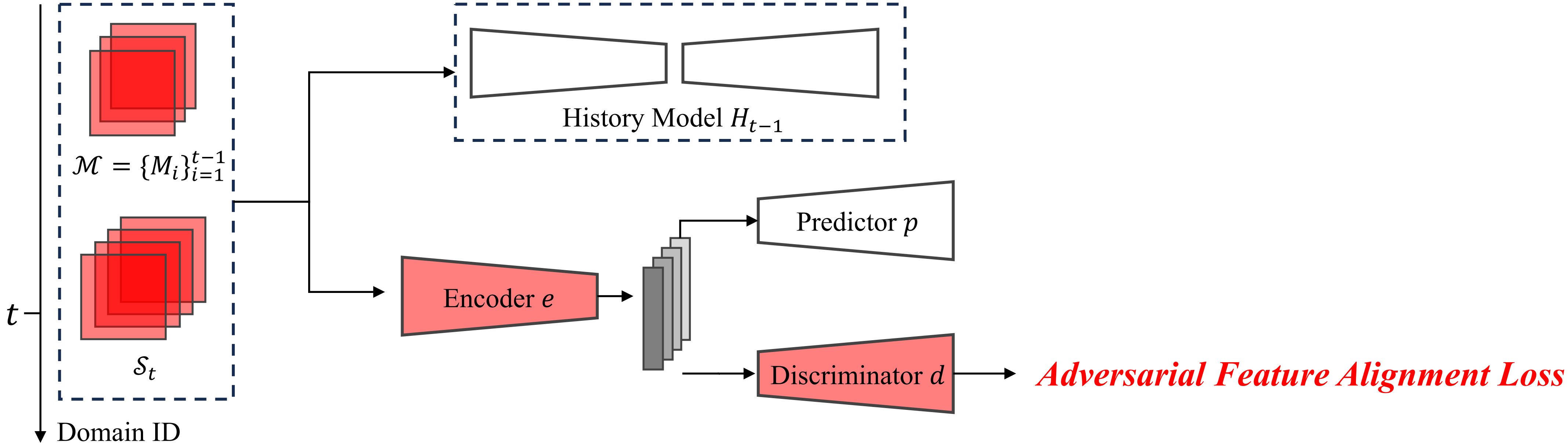


$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \le \left\{ \sum_{i=1}^{t-1} \left[\gamma_i \widehat{\epsilon}_{\mathcal{D}_i}(h) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_i}(h, H_{t-1})\right] \right\} + \left\{ \widehat{\epsilon}_{\mathcal{D}_t}(h) + \left(\sum_{i=1}^{t-1} \beta_i\right) \widehat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\}$$

$$+ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(H_{t-1})$$

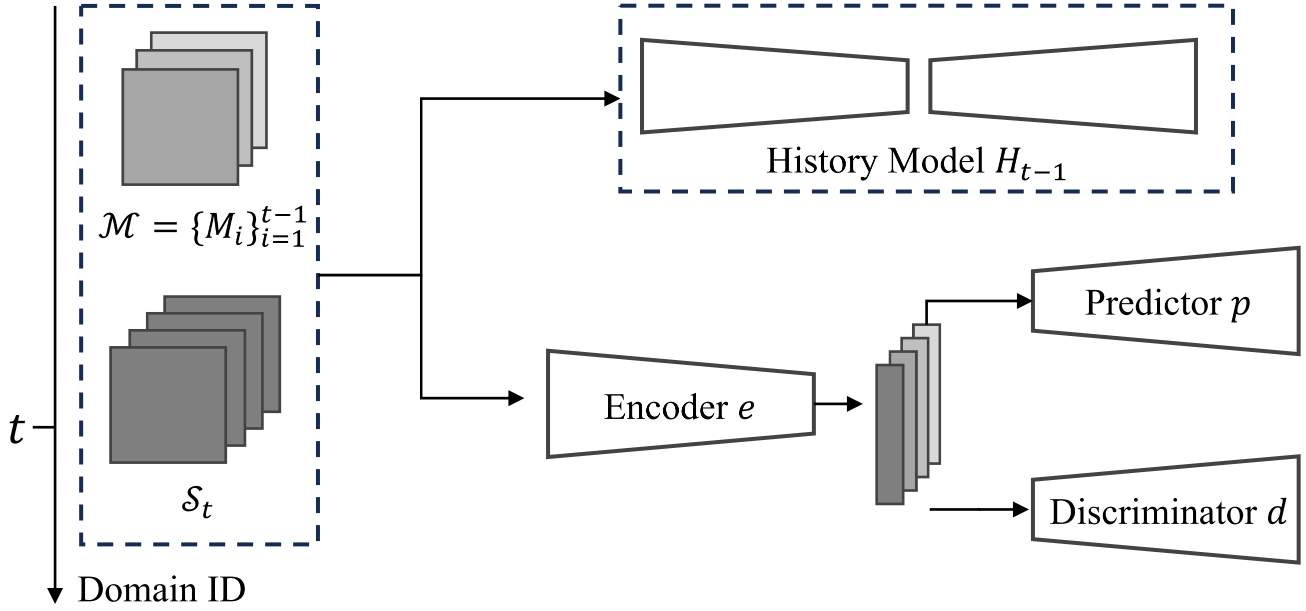
$$+ \sqrt{\left(\frac{(1+\sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i+\alpha_i)^2}{\widetilde{N}_i} \right) \left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}$$

UDIL: An *Adaptive* Bound for DIL



$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \widehat{\epsilon}_{\mathcal{D}_i}(h) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_i}(h, H_{t-1})] \right\} + \left\{ \widehat{\epsilon}_{\mathcal{D}_t}(h) + (\sum_{i=1}^{t-1} \beta_i) \widehat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\}$$
$$+ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(H_{t-1})$$
$$+ \sqrt{\left(\frac{(1+\sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\widetilde{N}_i} \right) \left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}$$

UDIL: An *Adaptive* Bound for DIL



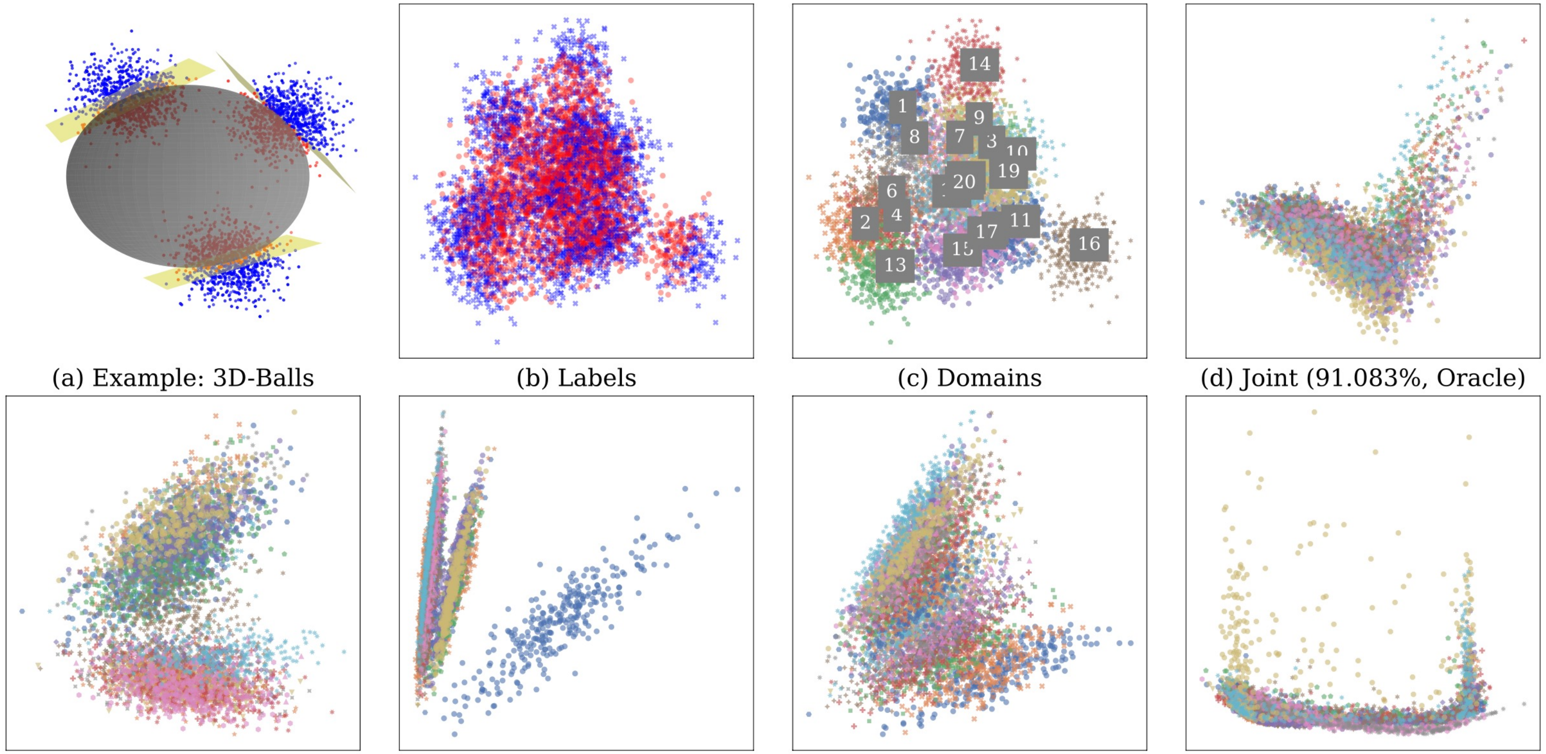
$\Omega = \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^{t-1}$

Adaptive Coefficient Optimization

$$\sum_{i=1}^{t} \epsilon_{\mathcal{D}_i}(h) \leq \left\{ \sum_{i=1}^{t-1} [\gamma_i \widehat{\epsilon}_{\mathcal{D}_i}(h) + \alpha_i \widehat{\epsilon}_{\mathcal{D}_i}(h, H_{t-1})] \right\} + \left\{ \widehat{\epsilon}_{\mathcal{D}_t}(h) + (\sum_{i=1}^{t-1} \beta_i) \widehat{\epsilon}_{\mathcal{D}_t}(h, H_{t-1}) \right\}$$
$$+ \frac{1}{2} \sum_{i=1}^{t-1} \beta_i d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \sum_{i=1}^{t-1} (\alpha_i + \beta_i) \epsilon_{\mathcal{D}_i}(H_{t-1})$$
$$+ \sqrt{\left(\frac{(1+\sum_{i=1}^{t-1} \beta_i)^2}{N_t} + \sum_{i=1}^{t-1} \frac{(\gamma_i + \alpha_i)^2}{\widetilde{N}_i} \right) \left(8d \log\left(\frac{2eN}{d}\right) + 8 \log\left(\frac{2}{\delta}\right)\right)}$$

UDIL: Experimental Results

- UDIL's representation distribution on synthetic dataset (high-dimensional balls)



(a) Example: 3D-Balls (b) Labels (c) Domains (d) Joint (91.083%, Oracle)

(e) ER (82.255%) (f) DER++ (79.322%) (g) CLS-ER (85.844%) (h) UDIL (86.872%, Ours)

UDIL: Experimental Results

RUTGERS UNIVERSITY

- UDIL evaluated on realistic datasets.

HD-Balls, Permuted-MNIST, Rotated-MNIST

Method	Buffer	HD-Balls Avg. Acc (↑)	HD-Balls Forgetting (↓)	P-MNIST Avg. Acc (↑)	P-MNIST Forgetting (↓)	R-MNIST Avg. Acc (↑)	R-MNIST Forgetting (↓)
Fine-tune	-	52.319±0.024	43.520±0.079	70.102±2.945	27.522±3.042	47.803±1.703	52.281±1.797
oEWC [47]	-	54.131±0.193	39.743±1.388	78.476±1.223	18.068±1.321	48.203±0.827	51.181±0.867
SI [60]	-	52.303±0.037	43.175±0.041	79.045±1.357	17.409±1.446	48.251±1.381	51.053±1.507
LwF [26]	-	51.523±0.065	25.155±0.264	73.545±2.646	24.556±2.789	54.709±0.515	45.473±0.565
GEM [31]	400	69.747±0.656	13.591±0.779	89.097±0.149	6.975±0.167	76.619±0.581	21.289±0.579
A-GEM [7]	400	62.777±0.295	12.878±1.588	87.560±0.087	8.577±0.053	59.654±0.122	39.196±0.171
ER [42]	400	82.255±1.552	9.524±1.655	88.339±0.044	7.180±0.029	76.794±0.696	20.696±0.744
DER++ [5]	400	79.332±1.347	13.762±1.514	**92.950±0.361**	3.378±0.245	84.258±0.544	13.692±0.560
CLS-ER [2]	400	85.844±0.165	5.297±0.281	91.598±0.117	3.795±0.144	81.771±0.354	15.455±0.356
ESM-ER [46]	400	71.995±3.833	13.245±5.397	89.829±0.698	6.888±0.738	82.192±0.164	16.195±0.150
UDIL (Ours)	400	**86.872±0.195**	**3.428±0.359**	92.666±0.108	**2.853±0.107**	**86.635±0.686**	**8.506±1.181**
Joint (Oracle)	∞	91.083±0.332	-	96.368±0.042	-	97.150±0.036	-

UDIL: Experimental Results

RUTGERS UNIVERSITY

- UDIL evaluated on realistic datasets.

Sequential CORe-50

Method	Buffer	$\mathcal{D}_{1:3}$	$\mathcal{D}_{4:6}$	$\mathcal{D}_{7:9}$	$\mathcal{D}_{10:11}$	Overall	
		Avg. Acc (↑)				Avg. Acc (↑)	Forgetting (↓)
Fine-tune	-	73.707±13.144	34.551±1.254	29.406±2.579	28.689±3.144	31.832±1.034	73.296±1.399
oEWC [51]	-	74.567±13.360	35.915±0.260	30.174±3.195	28.291±2.522	30.813±1.154	74.563±0.937
SI [66]	-	74.661±14.162	34.345±1.001	30.127±2.971	28.839±3.631	32.469±1.315	73.144±1.588
LwF [29]	-	80.383±10.190	28.357±1.143	31.386±0.787	28.711±2.981	31.692±0.768	72.990±1.350
GEM [34]	500	79.852±6.864	38.961±1.718	39.258±2.614	36.859±0.842	37.701±0.273	22.724±1.554
A-GEM [8]	500	80.348±9.394	41.472±3.394	43.213±1.542	39.181±3.999	43.181±2.025	33.775±3.003
ER [46]	500	90.838±2.177	79.343±2.699	68.151±0.226	65.034±1.571	66.605±0.214	32.750±0.455
DER++ [6]	500	92.444±1.764	88.652±1.854	80.391±0.107	78.038±0.591	78.629±0.753	21.910±1.094
CLS-ER [3]	500	89.834±1.323	78.909±1.724	70.591±0.322	⋆	⋆	⋆
ESM-ER [50]	500	84.905±6.471	51.905±3.257	53.815±1.770	50.178±2.574	52.751±1.296	25.444±0.580
UDIL (Ours)	500	**98.152±1.665**	**89.814±2.302**	**83.052±0.151**	**81.547±0.269**	**82.103±0.279**	**19.589±0.303**
GEM [34]	1000	78.717±4.831	43.269±3.419	40.908±2.200	40.408±1.168	41.576±1.599	18.537±1.237
A-GEM [8]	1000	78.917±8.984	41.172±4.293	44.576±1.701	38.960±3.867	42.827±1.659	33.800±1.847
ER [46]	1000	90.048±2.699	84.668±1.988	77.561±1.281	72.268±0.720	72.988±0.566	25.997±0.694
DER++ [6]	1000	89.510±5.726	92.492±0.902	88.883±0.794	86.108±0.284	86.392±0.714	13.128±0.474
CLS-ER [3]	1000	92.004±0.894	85.044±1.276	⋆	⋆	⋆	⋆
ESM-ER [50]	1000	85.120±4.339	54.852±5.511	61.714±1.840	55.098±3.834	58.932±0.959	20.134±0.643
UDIL (Ours)	1000	**98.648±1.174**	**93.447±1.111**	**90.545±0.705**	**87.923±0.232**	**88.155±0.445**	**12.882±0.460**
Joint (Oracle)	∞	-	-	-	-	99.137±0.049	-

Conclusion

- Proposed a principled framework, UDIL, for domain incremental learning with memory to ***unify various existing methods***.
- Theoretical analysis shows that different existing methods are equivalent to minimizing the same error bound with different ***fixed*** coefficients.
- UDIL allows ***adaptive*** coefficients during training, thereby always achieving the tightest bound and improving the performance.

Paper

Code