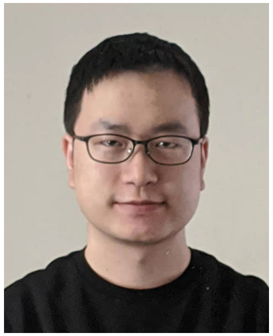
A Unified Meta-Learning Framework for Dynamic Transfer Learning



Jun Wu

UIUC

Jingrui He

UIUC

AIFARMS

Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

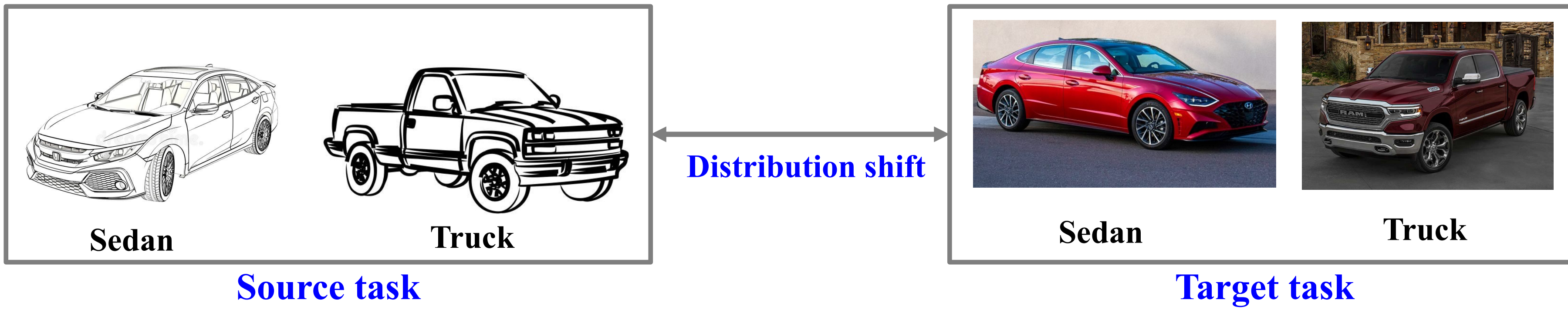
ILLINOIS

Roadmap

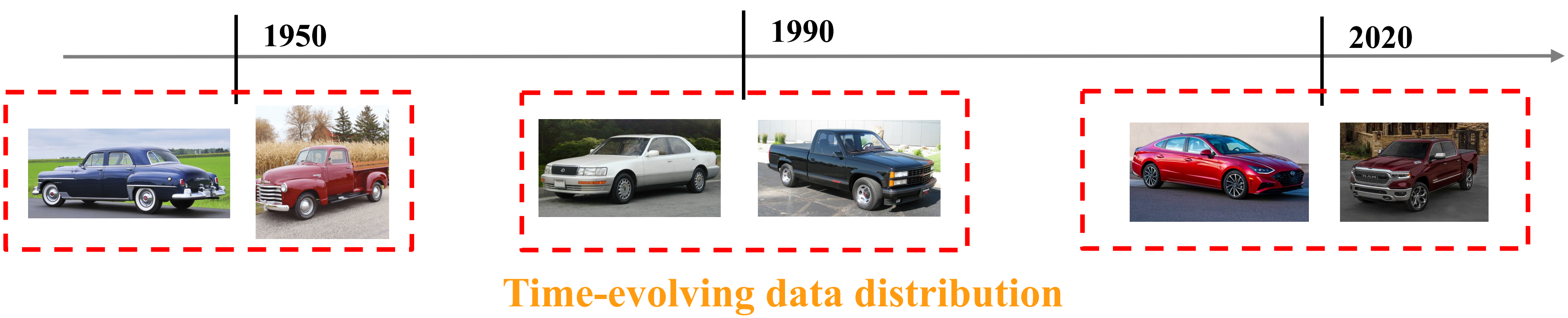
- Background
- Problem definition
- Theoretical analysis
- Proposed framework
- Experiments
- Conclusion

Background

- Distribution shift of transfer learning
  - E.g., sedans vs trucks



- Time-evolving distribution
  - New data are collected at different time stamps

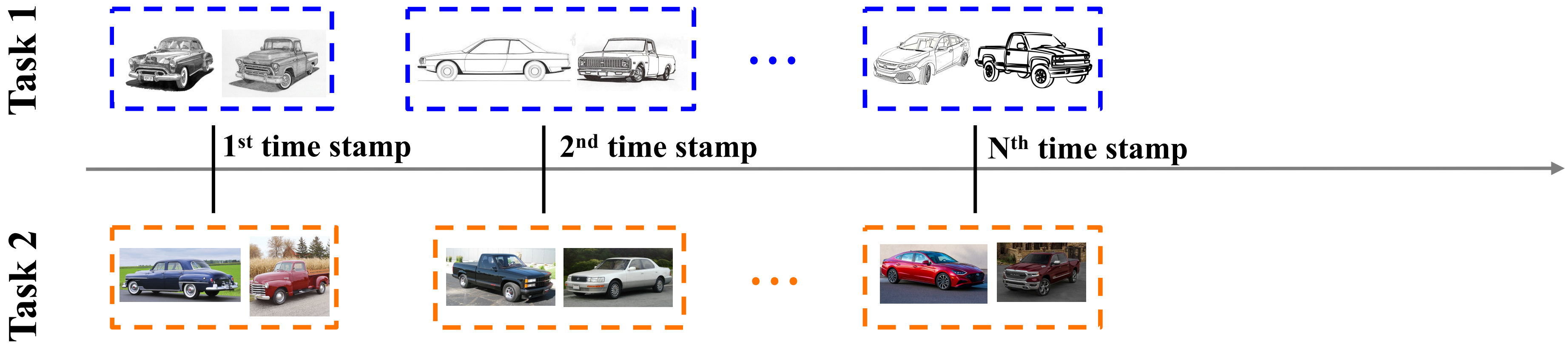


• Hoffman, Judy, Trevor Darrell, and Kate Saenko. "Continuous manifold based adaptation for evolving visual domains." CVPR. 2014

Problem Definition

- Dynamic transfer learning
  - Given: Labeled dynamic source task $\{\mathcal{D}_j^s\}_{j=1}^N$ (with data $D_j^s = \{x_{ij}^s, y_{ij}^s\}$);

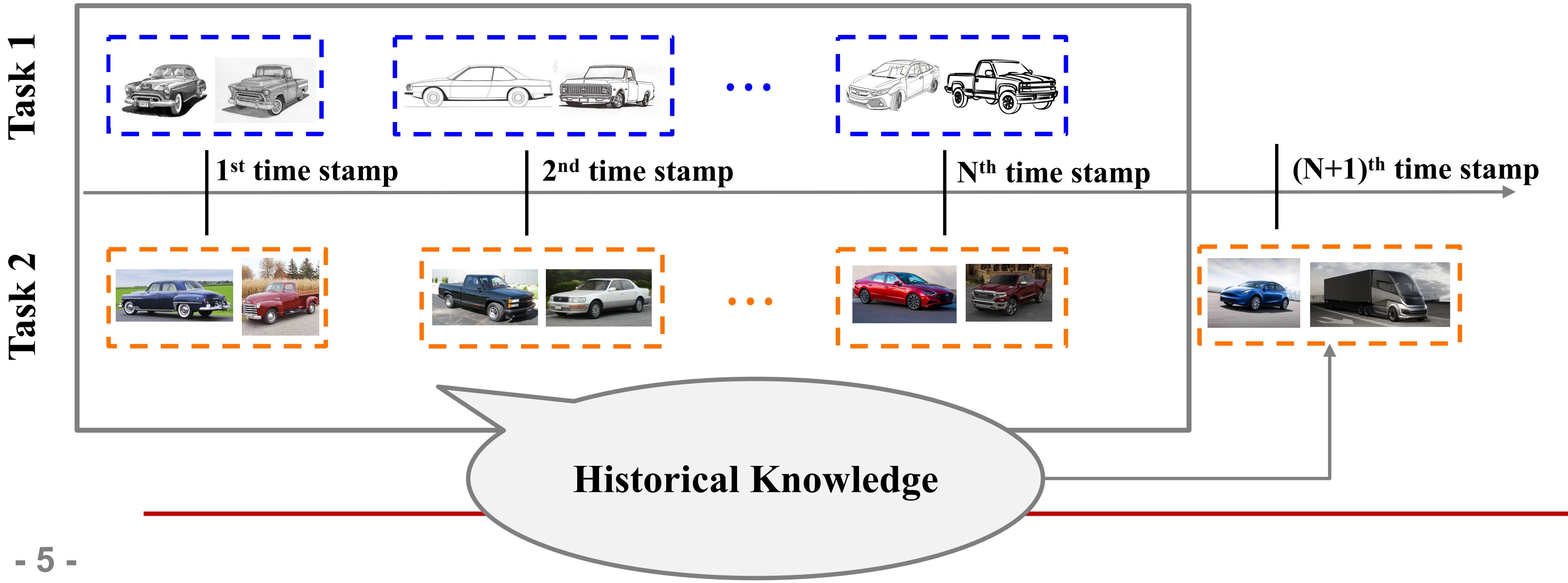
    unlabeled dynamic target task $\{\mathcal{D}_j^t\}_{j=1}^N$ (with data $D_j^t = \{x_{ij}^t\}$)



Problem Definition

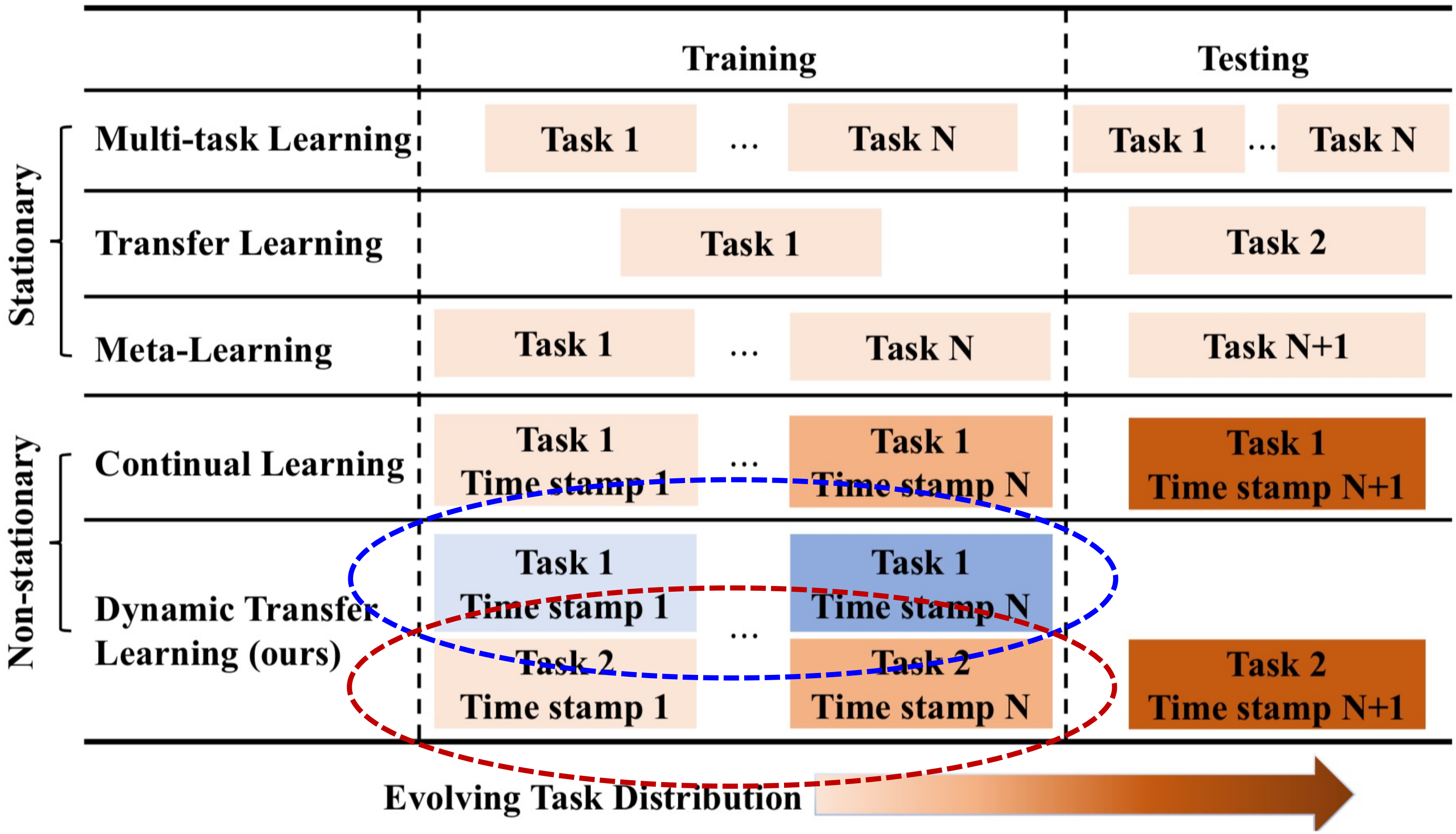
- Dynamic transfer learning
  - Given: Labeled dynamic source task $\{\mathcal{D}_j^s\}_{j=1}^N$ (with data $D_j^s = \{x_{ij}^s, y_{ij}^s\}$);

    unlabeled dynamic target task $\{\mathcal{D}_j^t\}_{j=1}^N$ (with data $D_j^t = \{x_{ij}^t\}$)
  - Goal: Learn the prediction function on the newest target task $\mathcal{D}_{N+1}^t$



Problem Comparison

- Dynamic transfer learning



- Sener, Ozan, and Vladlen Koltun. "Multi-task learning as multi-objective optimization." NeurIPS, 2018.
- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." ICML, 2017.
- Rolnick, David, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. "Experience replay for continual learning." NeurIPS, 2019.

Assumptions

- A1: The class labels of the source task are available at any time stamp
- A2: The source and target tasks are related at the initial time stamp
  - $d(\mathcal{D}_1^s, \mathcal{D}_1^t) \leq \Delta$ at time stamp $j = 1$
- A3: The data distributions of both source and target tasks are continuously changing over time
  - $d\left(\mathcal{D}_j^s, \mathcal{D}_{j+1}^s\right) \leq \Delta$ and $d\left(\mathcal{D}_j^t, \mathcal{D}_{j+1}^t\right) \leq \Delta$ for time stamp $j \geq 1$

Theoretical Analysis

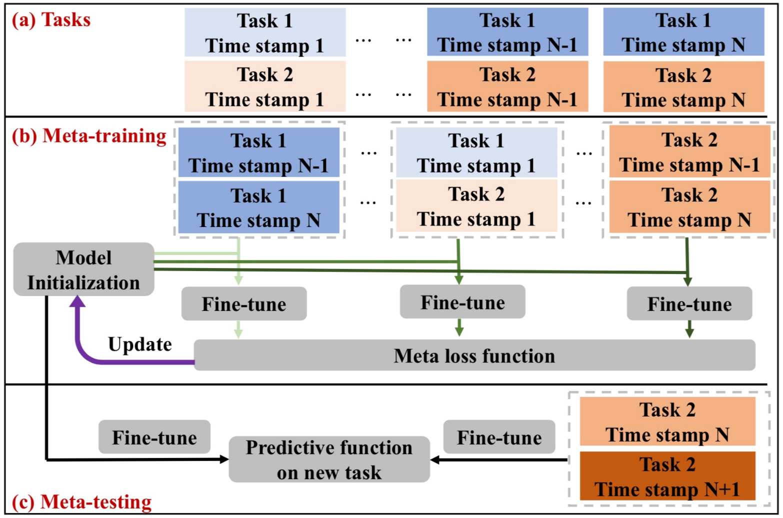
- Error bound on the newest target task $\mathcal{D}_{N+1}^{t}$
  - ① Empirical errors of historical source and target tasks;
  - ② Maximal distribution discrepancy across tasks and across time stamps;
  - ③ Maximal labeling difference across tasks and across time stamps;
  - ④ Average Rademacher complexity

Theorem 1: Assume that the loss function L is μ-admissible and obeys the triangle inequality, with probability at least $1-\delta$, the expected error ϵ_{N+1}^{t} for the newest target task $\mathcal{D}_{N+1}^{t}$ is bounded by

$$\epsilon_{N+1}^{t}(h) \leq \underbrace{\frac{1}{2N}\sum_{j=1}^{N}\left(\hat{\epsilon}_j^{s}(h)+\hat{\epsilon}_j^{t}(h)\right)}_{①} + \frac{N+2}{2}\left(\underbrace{d_{max}}_{②} + \underbrace{\lambda_{max}}_{③}\right) + \underbrace{\Re(H_L)}_{④} + \frac{\mu}{N}\sqrt{\frac{\log 1/\delta}{m}}$$

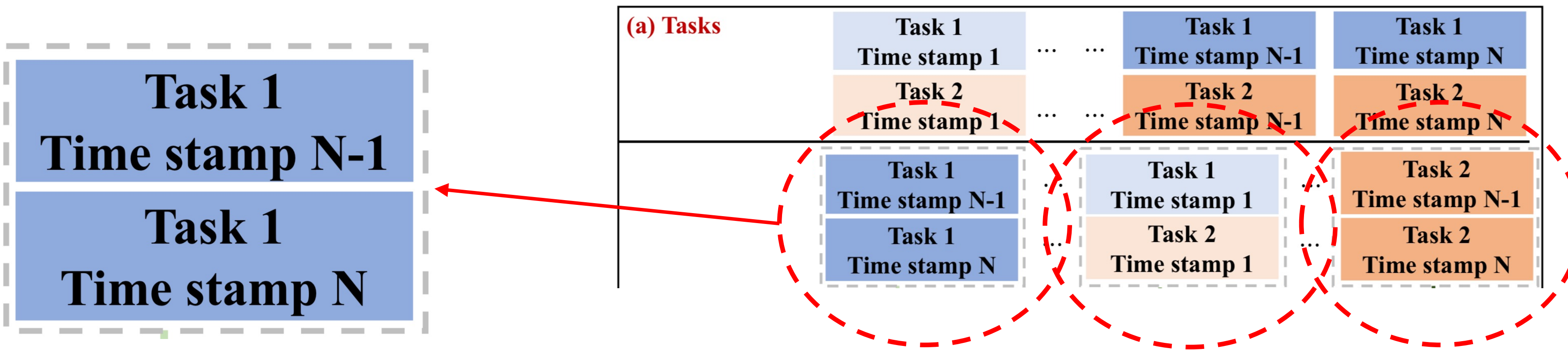
Proposed Framework: L2E

- A unified meta-learning framework
  - Learning to evolve (L2E)
- Three key components
  - Meta-pairs of tasks
  - Meta-training
  - Meta-testing



Step 1: Meta-pair of Tasks

- Construction of meta-pair of tasks
  - Consecutive source/target task
  - Initial source and target tasks



② Maximal distribution discrepancy across tasks and across time stamps;

$$d_{max} = \max\left\{\max_{1\le j\le N-1} d\left(\mathcal{D}_j^s, \mathcal{D}_{j+1}^s\right), d\left(\mathcal{D}_1^s, \mathcal{D}_1^t\right), \max_{1\le j\le N} d\left(\mathcal{D}_j^t, \mathcal{D}_{j+1}^t\right)\right\}$$

- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." ICML, 2017.

Step 2: Meta-training

- Objective function
  - Learn an optimal model initialization

$$\theta_N^* = \arg\min_{\theta} \sum_{k=1-N}^{N-1} \boxed{\zeta_k\left(M_k(\theta); D_k^{val}\right)}$$

Meta-pair of tasks

$$M_k(\theta) \leftarrow \theta - \alpha \cdot \nabla_\theta \zeta_k\left(\theta;\ D_k^{train}\right)$$

One-step GD

$$\zeta_k(\theta; D_k) = \underbrace{\begin{cases} \hat{\epsilon}_{-k+1}^s(\theta) + \gamma \cdot d(\mathcal{D}_{-k}^s, \mathcal{D}_{-k+1}^s); & k < 0 \\ \hat{\epsilon}_1^s(\theta) + \gamma \cdot d(\mathcal{D}_1^s, \mathcal{D}_1^t); & k = 0 \\ \hat{\epsilon}_k^t(\theta) + \gamma \cdot d(\mathcal{D}_k^t, \mathcal{D}_{k+1}^t); & k > 0 \end{cases}}_{\text{It is equivalent to standard transfer learning for each meta-pair of tasks}}$$

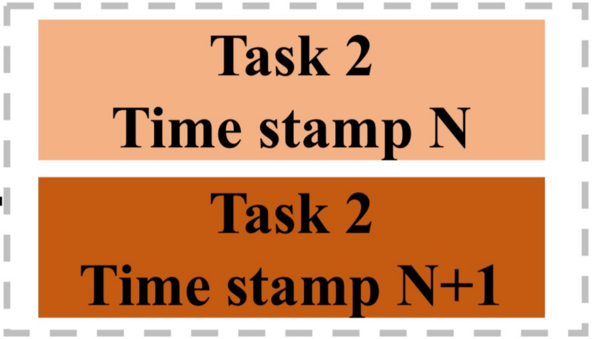
Meta-pairs of tasks

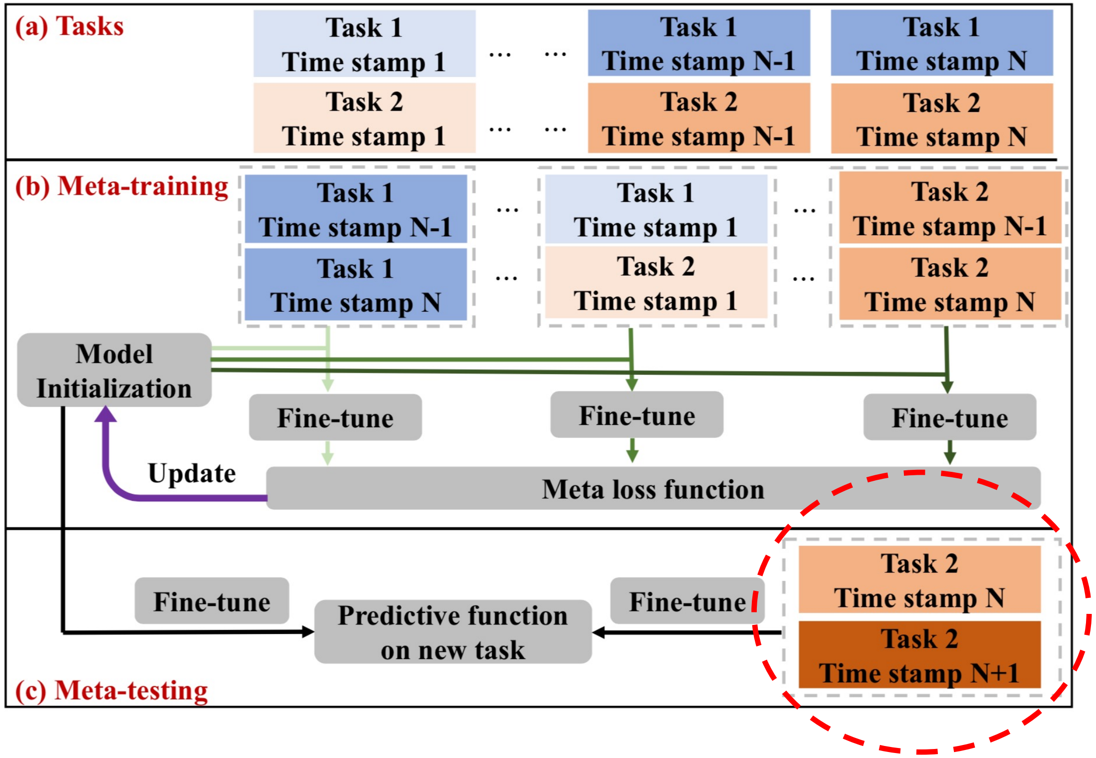
- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." ICML, 2017.

Step 3: Meta-testing

- Fine-tune θ_N^* for the newest target task

$$\theta_{N+1} = M_N(\theta_N^*) \leftarrow \theta_N^* - \alpha \cdot \nabla_\theta \zeta_N\left(\theta;\ D_N^{train}\right)$$





- Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." ICML, 2017.

Experiments

- Data sets
  - Office-31
  - Image-CLEF
  - Caltran

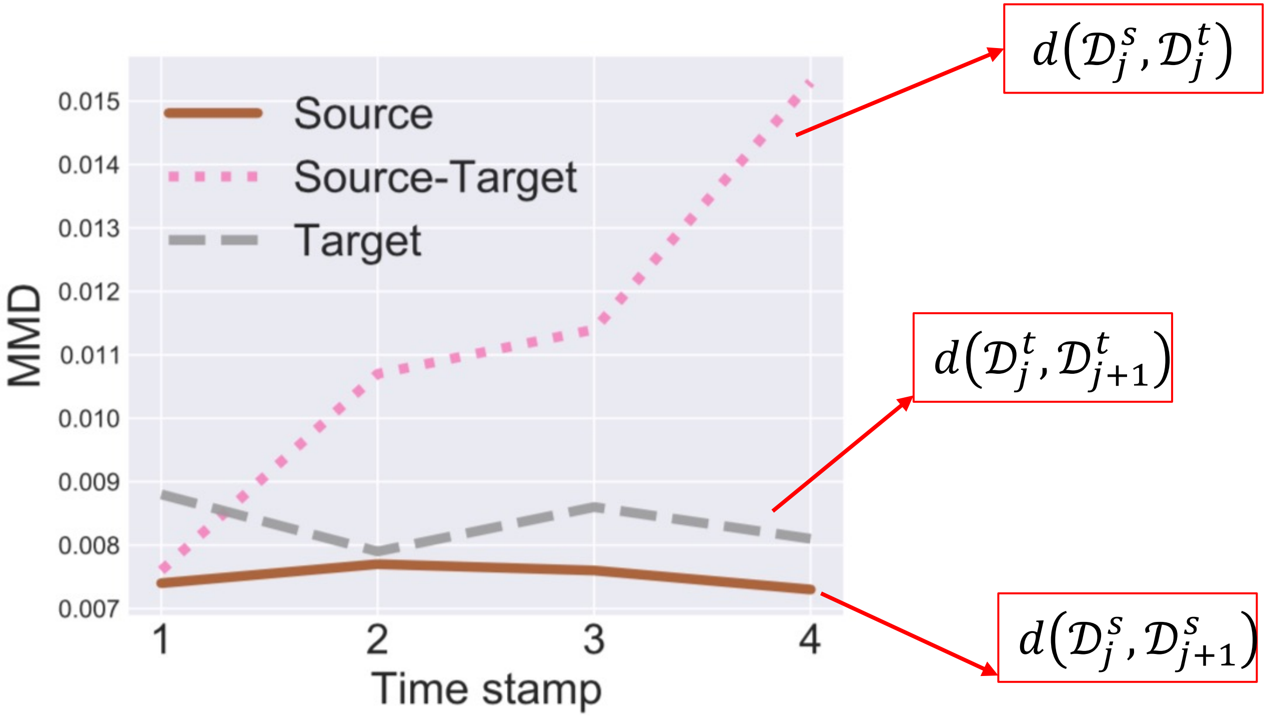
  Generate the dynamic task by adding the random noise and rotation to the original images

- Baselines
  - Static adaptation: SourceOnly, DANN, MDD
  - Multi-source adaptation: MDAN, M3SDA, DARN
  - Dynamic adaptation: CUA, TransLATE, GST
- Metric
  - Acc: Classification accuracy on the newest target task
  - H-Acc: Average classification accuracy on all the historical target tasks

- Zhang, Yuchen, et al. "Bridging theory and algorithm for domain adaptation." ICML. 2019.
- Zhao, Han, et al. "Adversarial multiple source domain adaptation." NeurIPS. 2018.
- Kumar, Ananya, Tengyu Ma, and Percy Liang. "Understanding self-training for gradual domain adaptation." ICML. 2020.

Evolution of Tasks

- Visualizing the distribution discrepancy on Image-CLEF (B→P)
  - MMD: Maximum mean discrepancy



(1) The source and target tasks are changing smoothly
(2) The relatedness between source and target tasks is decreasing over time

- Gretton, Arthur, et al. "A kernel two-sample test." *The Journal of Machine Learning Research* 13, no. 1 (2012): 723-773.

Results

	Method	I → C Acc	I → C H-Acc	B → P Acc	B → P H-Acc
Static adaptation	SourceOnly	$0.26_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.24_{\pm 0.00}$	$0.43_{\pm 0.01}$
	DANN	$0.36_{\pm 0.00}$	$0.58_{\pm 0.01}$	$0.27_{\pm 0.01}$	$0.43_{\pm 0.00}$
	MDD	$0.41_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.28_{\pm 0.03}$	$0.42_{\pm 0.01}$
Multi-source adaptation	MDAN	$0.62_{\pm 0.03}$	$0.77_{\pm 0.00}$	$0.37_{\pm 0.05}$	$0.51_{\pm 0.02}$
	M3SDA	$0.56_{\pm 0.03}$	$0.74_{\pm 0.01}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.02}$
	DARN	$0.55_{\pm 0.02}$	$0.76_{\pm 0.02}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.01}$
Dynamic adaptation	CUA	$0.58_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.36_{\pm 0.03}$	$0.51_{\pm 0.00}$
	TransLATE	$0.64_{\pm 0.01}$	$0.76_{\pm 0.00}$	$0.40_{\pm 0.03}$	$0.55_{\pm 0.01}$
	GST	$0.39_{\pm 0.01}$	$0.54_{\pm 0.03}$	$0.32_{\pm 0.01}$	$0.31_{\pm 0.02}$
	L2E (ours)	$\mathbf{0.66}_{\pm \mathbf{0.02}}$	$\mathbf{0.80}_{\pm \mathbf{0.01}}$	$\mathbf{0.44}_{\pm \mathbf{0.04}}$	$\mathbf{0.57}_{\pm \mathbf{0.02}}$

Results

Method	I → C		B → P	
	Acc	H-Acc	Acc	H-Acc
SourceOnly	$0.26_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.24_{\pm 0.00}$	$0.43_{\pm 0.01}$
DANN	$0.36_{\pm 0.00}$	$0.58_{\pm 0.01}$	$0.27_{\pm 0.01}$	$0.43_{\pm 0.00}$
MDD	$0.41_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.28_{\pm 0.03}$	$0.42_{\pm 0.01}$
MDAN	$0.62_{\pm 0.03}$	$0.77_{\pm 0.00}$	$0.37_{\pm 0.05}$	$0.51_{\pm 0.02}$
M3SDA	$0.56_{\pm 0.03}$	$0.74_{\pm 0.01}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.02}$
DARN	$0.55_{\pm 0.02}$	$0.76_{\pm 0.02}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.01}$
CUA	$0.58_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.36_{\pm 0.03}$	$0.51_{\pm 0.00}$
TransLATE	$0.64_{\pm 0.01}$	$0.76_{\pm 0.00}$	$0.40_{\pm 0.03}$	$0.55_{\pm 0.01}$
GST	$0.39_{\pm 0.01}$	$0.54_{\pm 0.03}$	$0.32_{\pm 0.01}$	$0.31_{\pm 0.02}$
L2E (ours)	$\mathbf{0.66}_{\pm \mathbf{0.02}}$	$\mathbf{0.80}_{\pm \mathbf{0.01}}$	$\mathbf{0.44}_{\pm \mathbf{0.04}}$	$\mathbf{0.57}_{\pm \mathbf{0.02}}$

Static adaptation: SourceOnly, DANN, MDD

Multi-source adaptation: MDAN, M3SDA, DARN

Dynamic adaptation: CUA, TransLATE, GST

(1) Higher performance on the newest target task

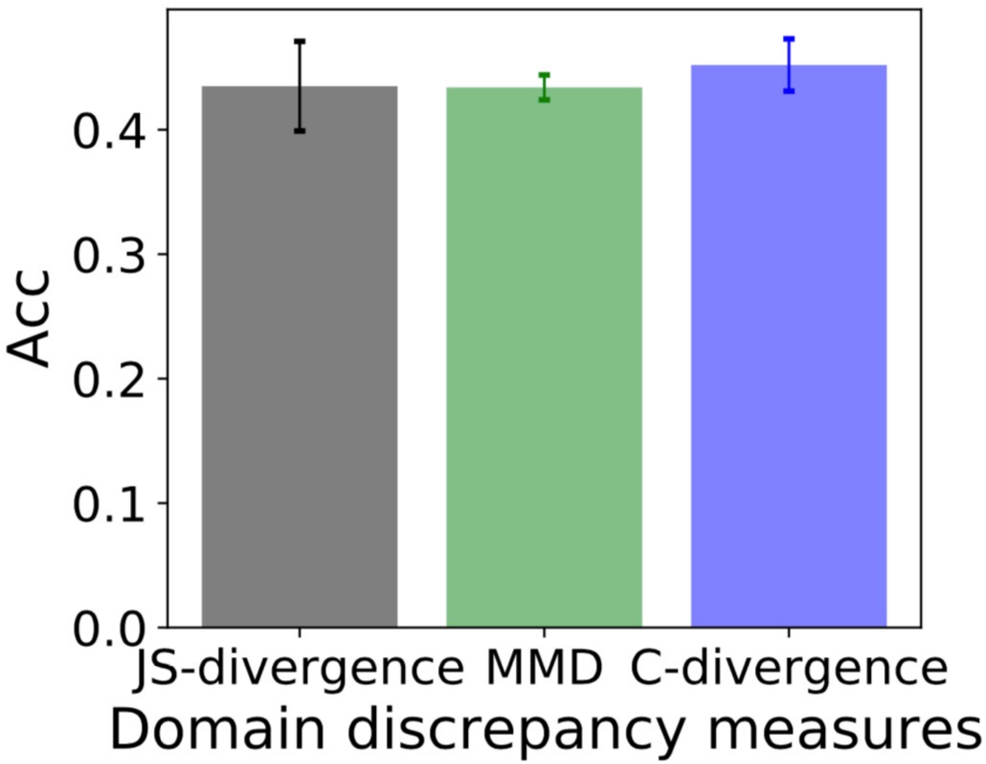
Results

	Method	I → C Acc	I → C H-Acc	B → P Acc	B → P H-Acc
Static adaptation	SourceOnly	$0.26_{\pm 0.01}$	$0.51_{\pm 0.01}$	$0.24_{\pm 0.00}$	$0.43_{\pm 0.01}$
	DANN	$0.36_{\pm 0.00}$	$0.58_{\pm 0.01}$	$0.27_{\pm 0.01}$	$0.43_{\pm 0.00}$
	MDD	$0.41_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.28_{\pm 0.03}$	$0.42_{\pm 0.01}$
Multi-source adaptation	MDAN	$0.62_{\pm 0.03}$	$0.77_{\pm 0.00}$	$0.37_{\pm 0.05}$	$0.51_{\pm 0.02}$
	M3SDA	$0.56_{\pm 0.03}$	$0.74_{\pm 0.01}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.02}$
	DARN	$0.55_{\pm 0.02}$	$0.76_{\pm 0.02}$	$0.39_{\pm 0.02}$	$0.52_{\pm 0.01}$
Dynamic adaptation	CUA	$0.58_{\pm 0.01}$	$0.74_{\pm 0.01}$	$0.36_{\pm 0.03}$	$0.51_{\pm 0.00}$
	TransLATE	$0.64_{\pm 0.01}$	$0.76_{\pm 0.00}$	$0.40_{\pm 0.03}$	$0.55_{\pm 0.01}$
	GST	$0.39_{\pm 0.01}$	$0.54_{\pm 0.03}$	$0.32_{\pm 0.01}$	$0.31_{\pm 0.02}$
	L2E (ours)	$\mathbf{0.66}_{\pm \mathbf{0.02}}$	$\mathbf{0.80}_{\pm \mathbf{0.01}}$	$\mathbf{0.44}_{\pm \mathbf{0.04}}$	$\mathbf{0.57}_{\pm \mathbf{0.02}}$

(1) Higher performance on the newest target task
(2) Higher performance on the historical target task

Ablation Study

- Flexibility: L2E with different distribution discrepancy measures



- Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The Journal of Machine Learning Research*17, no. 1 (2016): 2096-2030.
- Gretton, Arthur, et al. "A kernel two-sample test." *The Journal of Machine Learning Research* 13, no. 1 (2012): 723-773.
- Wu, Jun, et al. "Continuous transfer learning with label-informed distribution alignment." *arXiv preprint arXiv:2006.03230* (2020).

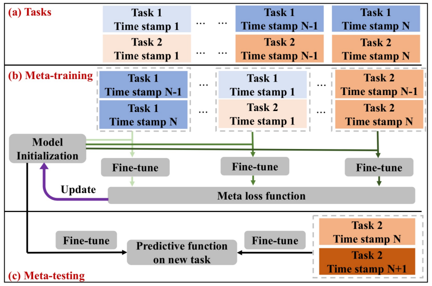
Conclusion

- Theoretical results
  - Derive the **generalization error bounds** of dynamic transfer learning

Theorem 1: Assume that the loss function L is μ-admissible and obeys the triangle inequality, with probability at least $1-\delta$, the expected error ϵ_{N+1}^{t} for the newest target task $\mathcal{D}_{N+1}^{t}$ is bounded by

$$\epsilon_{N+1}^{t}(h) \leq \frac{1}{2N}\sum_{j=1}^{N}\left(\hat{\epsilon}_j^{s}(h)+\hat{\epsilon}_j^{t}(h)\right)+\frac{N+2}{2}(d_{max}+\lambda_{max})+\Re(H_L)+\frac{\mu}{N}\sqrt{\frac{\log 1/\delta}{m}}$$

- Generic framework
  - Propose a **meta-learning framework** (L2E) by reformulating the meta-pairs of tasks



- Empirical evaluation
  - Demonstrate the **effectiveness** of our L2E framework on dynamic tasks

Method	I → C		B → P	
	Acc	H-Acc	Acc	H-Acc
SourceOnly	0.26±0.01	0.51±0.01	0.24±0.00	0.43±0.01
DANN	0.36±0.00	0.58±0.01	0.27±0.01	0.43±0.00
MDD	0.41±0.01	0.62±0.01	0.28±0.03	0.42±0.01
MDAN	0.62±0.03	0.77±0.00	0.37±0.05	0.51±0.02
M3SDA	0.56±0.03	0.74±0.01	0.39±0.02	0.52±0.02
DARN	0.55±0.02	0.76±0.02	0.39±0.02	0.52±0.01
CUA	0.58±0.01	0.74±0.01	0.36±0.03	0.51±0.00
TransLATE	0.64±0.01	0.76±0.00	0.40±0.03	0.55±0.01
GST	0.39±0.01	0.54±0.03	0.32±0.01	0.31±0.02
L2E (ours)	**0.66**±0.02	**0.80**±0.01	**0.44**±0.04	**0.57**±0.02

Thank You!

Please email me via **junwu3@illinois.edu** if you have any question.