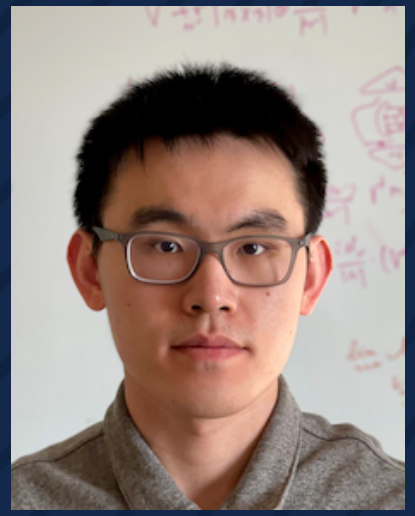
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Understanding Gradual Domain Adaptation:
Improved Analysis, Optimal Path and Beyond

ICML 2022



Haoxiang Wang
PhD Candidate
ECE, UIUC

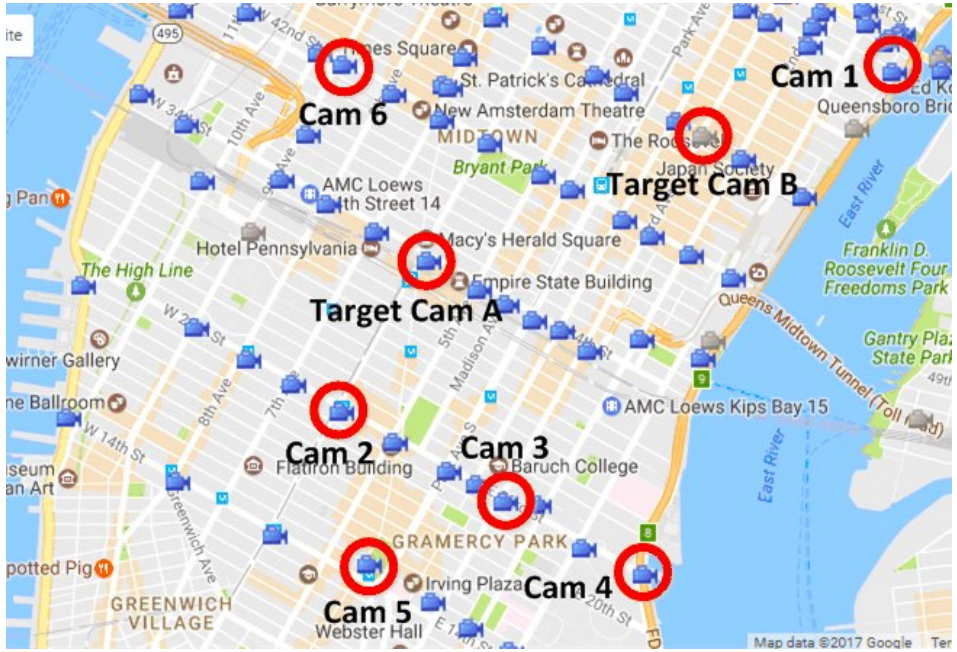
Bo Li
Assistant Professor
Computer Science, UIUC

Han Zhao
Assistant Professor
Computer Science, UIUC

Domain Adaptation: Source Domain ⟶ Target Domain

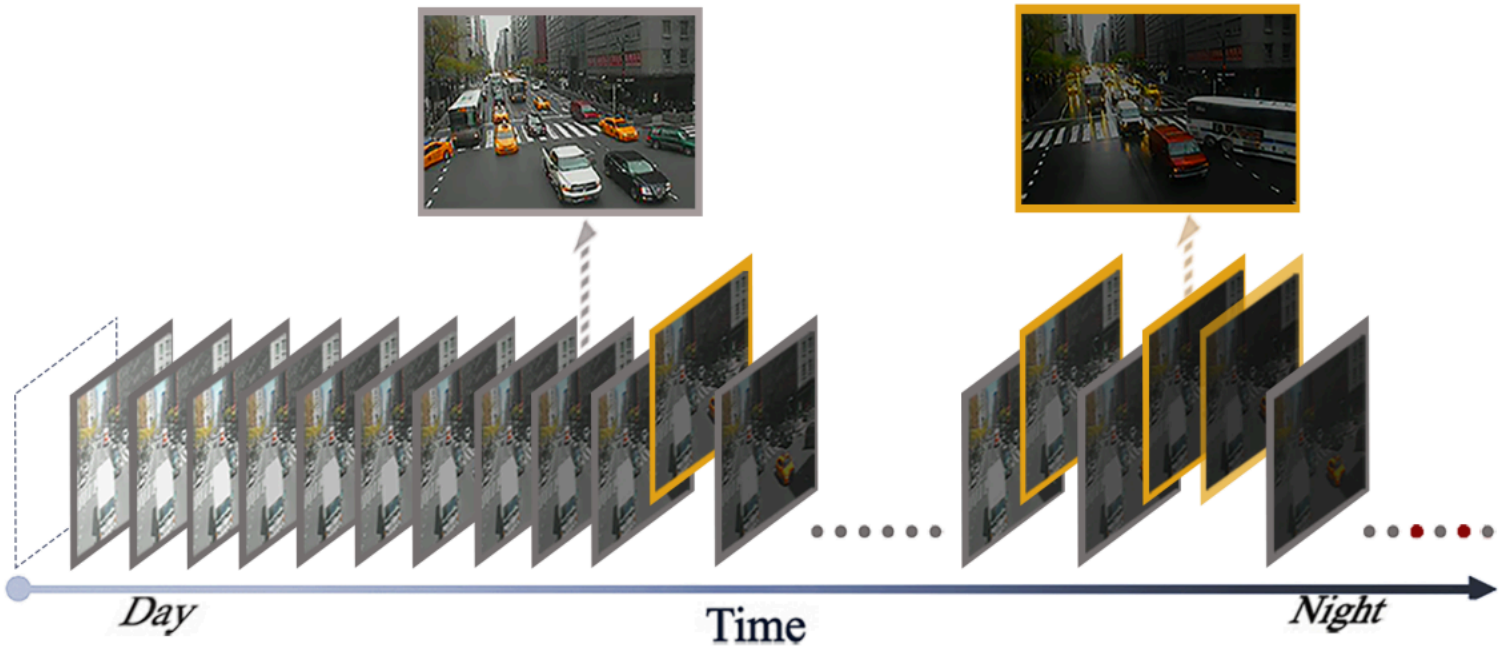
Source (with Labels)

Target (No Labels)

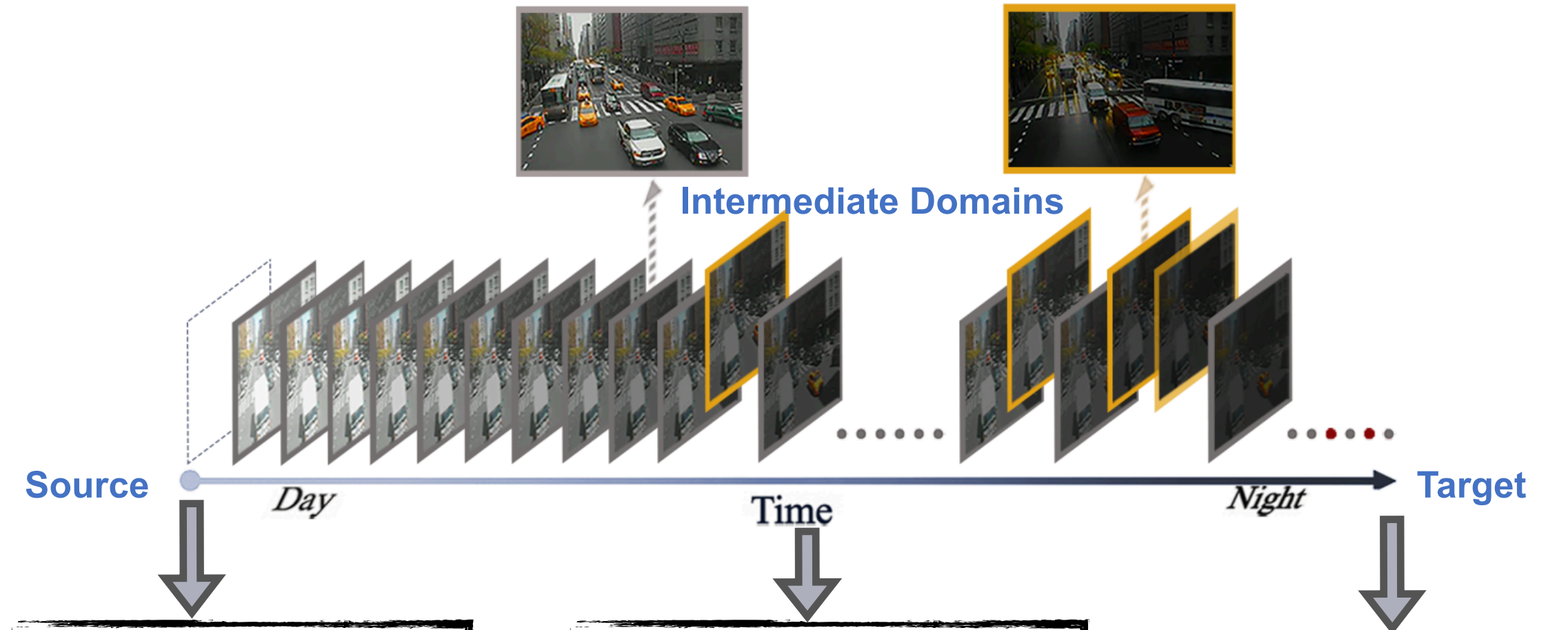


Gradual Domain Adaptation: Source ⟶ Intermediate Domains ⟶ Target

- In many real-world applications (e.g., self-driving cars, recommendation systems), distributions do not change abruptly
- Instead, they change **smoothly** in time / space
- Could we do better in this case?



Gradual Self-Training [Kumar, Ma and Liang. ICML 2020]



Existing Theory of Gradual Self-training

With high probability,

$$\varepsilon_T(h_T) \leq e^{\mathcal{O}(T)}\left(\varepsilon_0(h_0)+\mathcal{O}\left(\frac{1}{\sqrt{n}}+\sqrt{\frac{\log T}{n}}\right)\right)$$

- h_0: initial classifier; h_T: classifier evaluated at the target domain
- ε_0: error in the source domain; ε_T: error in the target domain
- T: # of intermediate domains (time steps) - 1
- n: # of unlabeled data points in each intermediate/target domain

[Kumar, Ma and Liang. *Understanding Self-Training for Gradual Domain Adaptation*. ICML 2020]

Existing Theory of Gradual Self-training

With high probability,

$$\varepsilon_T(h_T) \leq e^{\mathcal{O}(T)}\left(\varepsilon_0(h_0)+\mathcal{O}\left(\frac{1}{\sqrt{n}}+\sqrt{\frac{\log T}{n}}\right)\right)$$

The dependency on T is an **exponential** (pessimistic)

Is it possible to have a more optimistic generalization bound of the gradual self-training algorithm (hopefully with some insights for algorithm design)?

[Kumar, Ma and Liang. *Understanding Self-Training for Gradual Domain Adaptation*. ICML 2020]

Our Improved Analysis of Gradual Self-training

Theorem 1 (Generalization Bound for Gradual Self-Training). *For any $\delta \in (0, 1)$, the population loss of gradually self-trained classifier h_T in the target domain is upper bounded with probability at least $1 - \delta$ as*

$$\varepsilon_T(h_T) \leq \varepsilon_0(h_0) + \mathcal{O}\left(T\Delta + \frac{T}{\sqrt{n}} + T\sqrt{\frac{\log 1/\delta}{n}} + \frac{1}{\sqrt{nT}} + \sqrt{\frac{(\log nT)^{3L-2}}{nT}} + \sqrt{\frac{\log 1/\delta}{nT}}\right)$$

- h_0: initial classifier; h_T: classifier evaluated at the target domain
- T: # of intermediate domains (time steps) - 1
- n: # of unlabeled data points in each domain
- Δ: average distributional distance between consecutive domains

Proof Sketch

1. If the distribution distance is small, then for any smooth model, the error difference is also small.

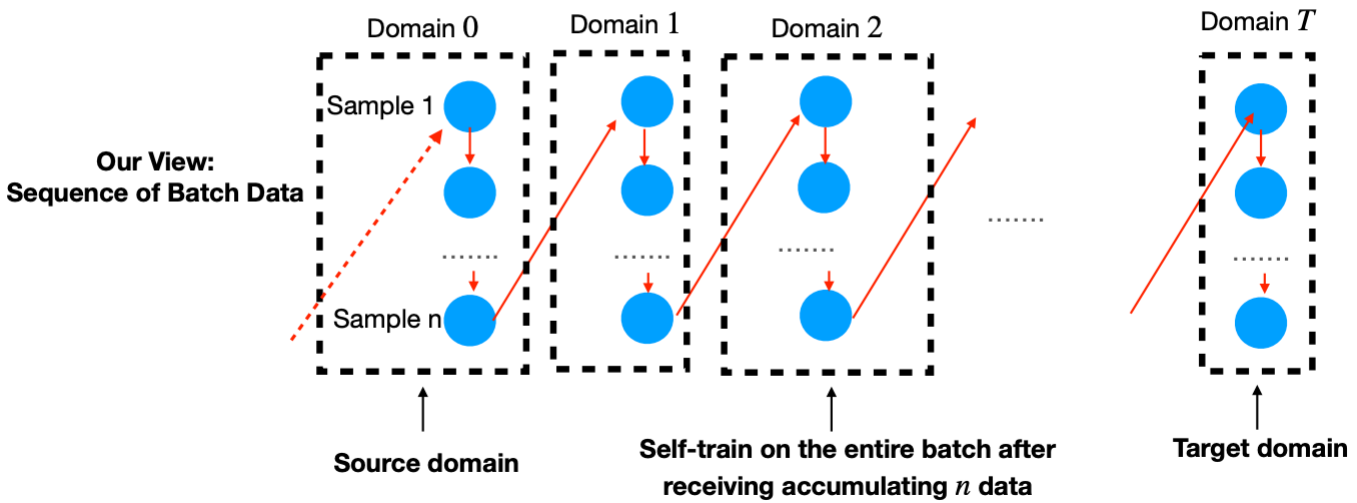
Lipschitz constants *p-Wasserstein Distance*

$$|\varepsilon_\mu(h) - \varepsilon_\nu(h)| \leq \rho\sqrt{R^2 + 1} W_p(\mu, \nu)$$

2. The self-training algorithm is stable:

$$\hat{h} = \arg\min_{f \in \mathcal{H}} \sum_{x \in S} \ell(f(x), h(x)) \qquad \text{s.t.} \qquad \left|\varepsilon_\mu(\hat{h}) - \varepsilon_\nu(h)\right| \leq \mathcal{O}\left(W_p(\mu, \nu) + \frac{\rho B + \sqrt{\log \frac{1}{\delta}}}{\sqrt{n}}\right)$$

3. A reduction to online learning (treating the pseudo-label as if they were ground-truth)



Comparison of Generation Bounds for Gradual Self-Training

The dependency on T is an **exponential** (pessimistic)

[Kumar, Ma, Liang. ICML'20]: $\varepsilon_T(h_T) \leq e^{\mathcal{O}(T)}\left(\varepsilon_0(h_0) + \mathcal{O}\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log T}{n}}\right)\right)$

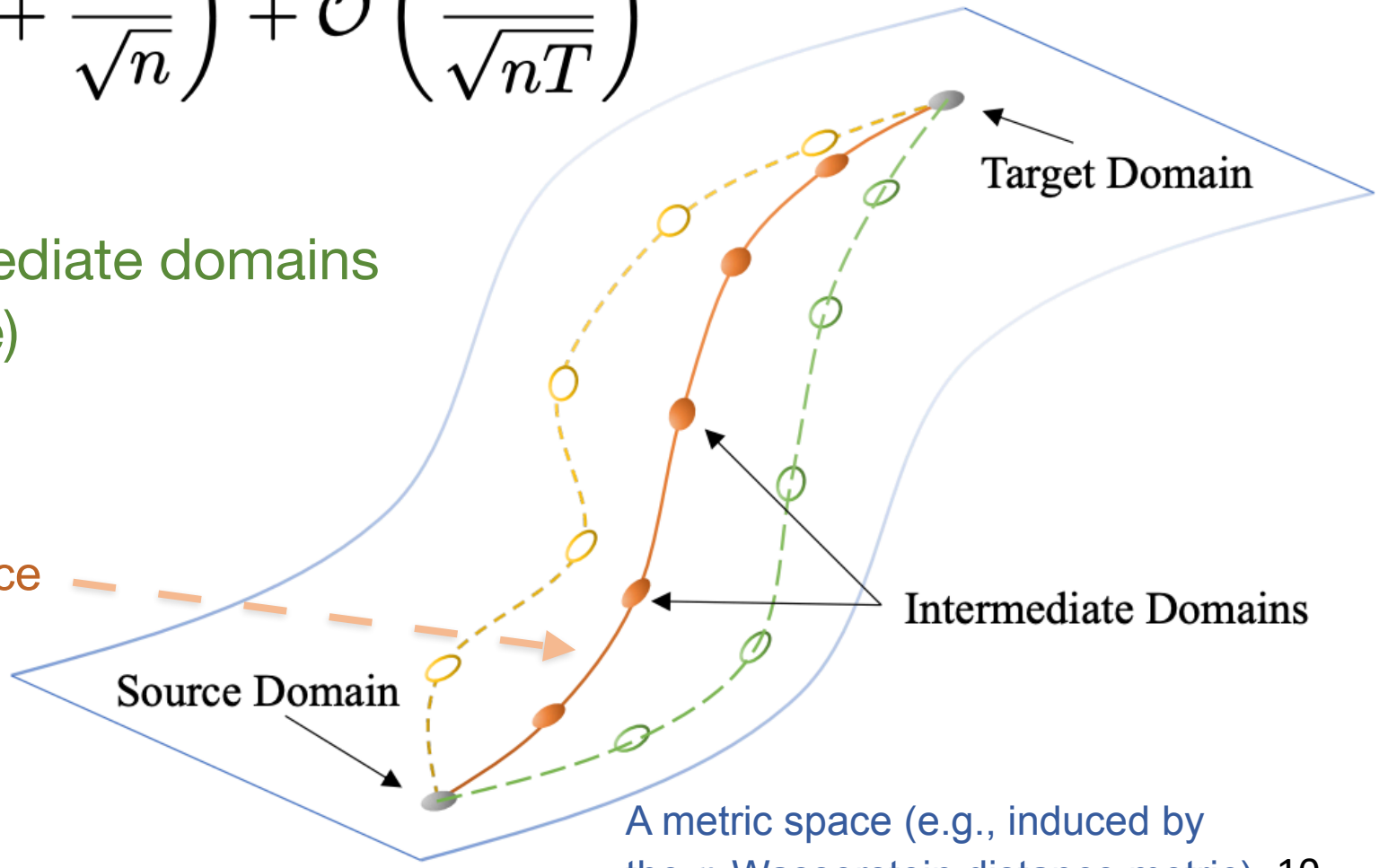
Ours: $\varepsilon_T(h_T) \leq \varepsilon_0(h_0) + \mathcal{O}\left(T\Delta + \frac{T}{\sqrt{n}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{nT}}\right)$

The dependency on T becomes **linear** (optimistic)

Optimal Path of Intermediate Domains

$$\varepsilon_T(h_T) \leq \varepsilon_0(h_0) + \mathcal{O}\left(T\Delta + \frac{T}{\sqrt{n}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{nT}}\right)$$

$T\Delta$: length of the path of intermediate domains (measured in some metric space)



Optimal Number of Intermediate Domains

$$\varepsilon_T(h_T) \leq \varepsilon_0(h_0) + \mathcal{O}\left(T\Delta + \frac{T}{\sqrt{n}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{nT}}\right)$$

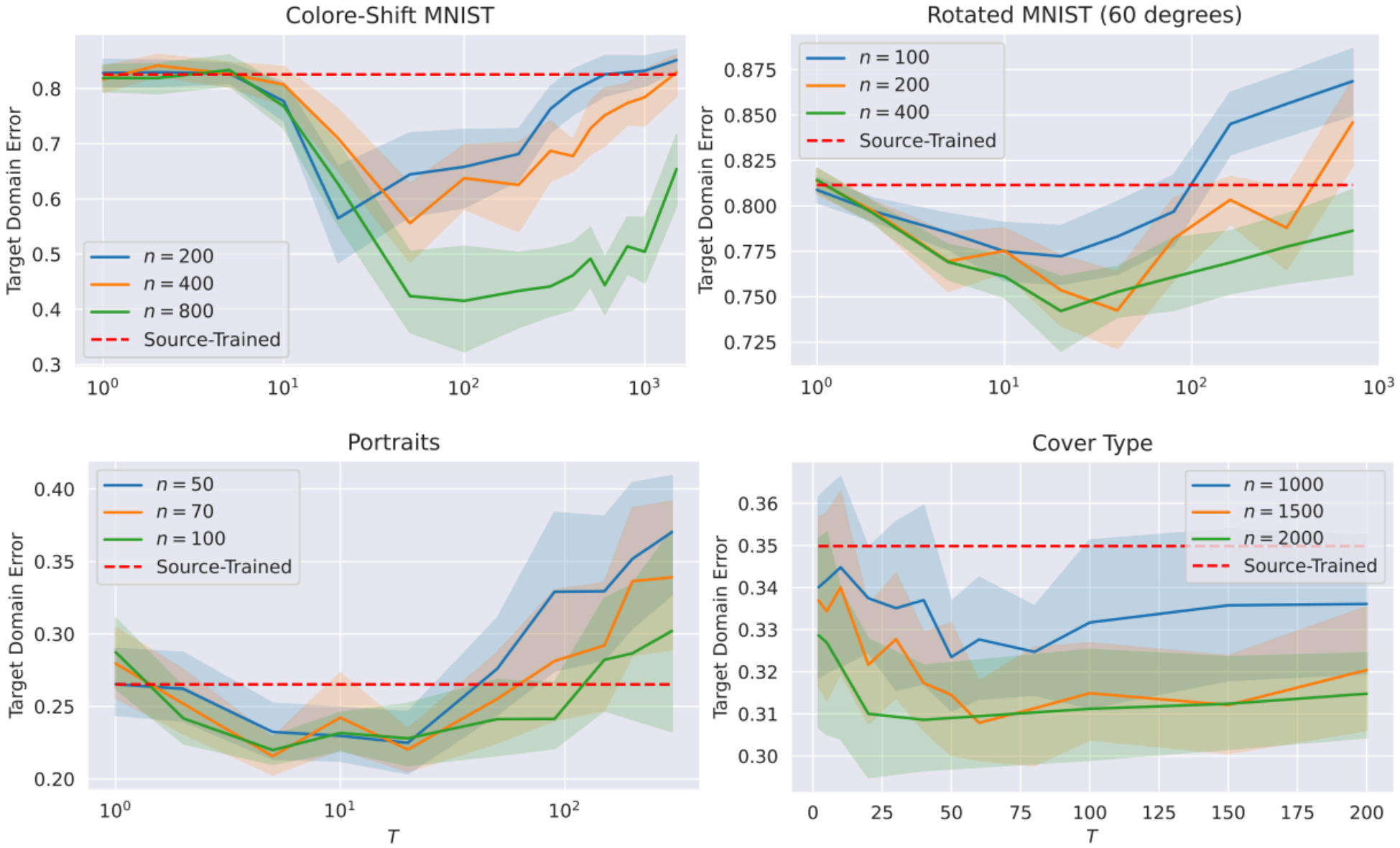
Minimize our error bound w.r.t. T

$$T^* = \max\left\{\frac{L}{\Delta_{\max}}, \tilde{\mathcal{O}}\left(\left(\frac{1}{1+\Delta_{\max}\sqrt{n}}\right)^{2/3}\right)\right\}$$

- $\Delta_{\max}$: average distributional distance between consecutive domains

Is there an optimal T^* in practice?

$$T^* = \max\left\{\frac{L}{\Delta_{\max}}, \tilde{\mathcal{O}}\left(\left(\frac{1}{1+\Delta_{\max}\sqrt{n}}\right)^{2/3}\right)\right\}$$



Thank you for watching this presentation!

Code: github.com/Haoxiang-Wang/gradual-domain-adaptation

Contact Information:

- Haoxiang Wang: hwang264@illinois.edu
- Bo Li: lbo@illinois.edu
- Han Zhao: hanzhao@illinois.edu