Many thanks to all reviewers for your constructive and insightful comments. We address your concerns as follows.

Response to common questions

Table 1: Comparison of settings in related work.

Method	Online	Evolvement	Small-Batch	Forgetting
CMA [10] and IEDA	✓	✓	✓	×
Bobu *et al.* [2]	×	×	×	✓
Lao *et al.* [14]	✓	×	×	✓
EAML	✓	✓	✓	✓

Q1: More comparison with other continual/incremental learning literature.
A1: We compared with *DANN+Replay* as a continual learning method (in Table 1 of the main text). We add *DANN+EWC* and *DANN+GEM* in Table 3. We will elaborate on continual/incremental learning literature in the revision.

Q2: Comparison with other evolving/continual domain methods including [2], [10], [14].
A2: Existing continuous or evolving domain refers to **different settings** though in the name of "continuous DA", thus they cannot be adapted to our settings directly. See the *comparison of these settings* in Table 1. Nonetheless, we adapt *CMA [10]* and *Incremental Evolving Domain Adaptation (IEDA)* as evolving domain methods and provide results in Table 3. Due to the missing details and codes of the preprint of [14], we follow their idea and design *DANN+Replay* (Replay Oracle of Table 1 in the main text). We will discuss these settings and methods in detail.

Response to Reviewer 6

Table 2: Accuracy (%) on Vehicles.

Method	1995	2000	2005	2010	2015	2020
DANN	51.0±1.5	54.9±0.8	58.6±1.2	65.5±1.0	69.3±1.1	70.4±1.5
JAN Merge	67.2±1.8	68.4±1.6	66.0±0.9	65.7±1.0	64.9±1.4	64.3±1.0
EAML	**70.1**±1.3	**69.8**±0.9	**72.5**±0.8	**73.6**±1.3	**75.5**±1.4	**75.2**±1.1

Q1: Missing quantified results on Caltran and Vehicles.
A1: In the main text, we provided results on Caltran and Vehicles in figures due to limitation of space. We further provide numerical results in Table 2 and will add it to the next version.

Q2: MNIST and Vehicles are not continuous enough.
A2: On MNIST and Vehicles, we sample *random* angles in $0° - 60°$ and years in $1980 - 1995$ during meta-training so that they are *sufficiently continuous*. In meta-testing, we use the same target trajectory on each method for fair and convenient comparison and set the trajectory to $\{120°, 126° \cdots 180°\}$ in MNIST and $\{2000, 2005, 2010, 2015\}$ in Vehicles. To further address the concern, we provide the *randomly sampled meta-testing results* in Table 3.

Table 3: Rotated MNIST $120° - 180°$ random test.

Method	Average acc.
Source Only	27.15
DANN	29.40
CMA [10]	30.65
IEDA	30.81
MAML	31.34
JAN-Merge offline	31.58
DANN + EWC	32.66
DANN + GEM	33.40
EAML	**35.03**

Response to Reviewer 7

Q1: It is not clear how to balance domain adaptation and learning without forgetting.
A1: The balance of adaptation and avoiding forgetting relies on a *hyper-parameter* λ as in Line 7 of Algorithm 1 in main text and *the learning rate of meta-adapters* $\eta_{\phi'}$. Larger λ and $\eta_{\phi'}$ indicate larger penalty on forgetting. We provide *performance on Caltran* w.r.t λ and $\eta_{\phi'}$ in Figure 1.

Q2: Why does JAN-Merge underperform EAML-full and EAML-rep at the beginning?
A2: JAN-Merge merges all target data and adapts to them. Without capturing the evolvement of target data, adapting to them as a whole results in *conflict among evolvement* and hurts the performance on the early-adapted target domains. In Figure 3(b) of main text, EAML performs similarly to DANN at the beginning, but JAN-Merge underperforms EAML and DANN.

Q4: Comparison with "ACE: Adapting to Changing Environments for Semantic Segmentation".
A4: This reference deals with segmentation, which is a different setting and cannot be compared with our work directly. We will discuss it in detail in a future version.

Q5: How to address large intra-variance within target domain and abrupt change?
A5: α in Line 83 of main text describes the rate of change in the target. As Theorem 1 points out, the learner generalizes to held-out target data in meta-testing only for reasonably small α, which our work falls under. Sudden change in target data is an interesting topic for future work.

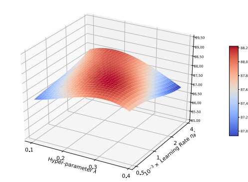

Figure 1: Accuracy on Caltran w.r.t λ and $\eta_{\phi'}$.

Response to Reviewer 8

Q1: The model does not consistently perform better than baselines for different settings.
A1: EAML does outperform *all baselines on the average of evolving target data*. For each rotation of MNIST, the performance can be *affected by the sampling of test samples*. Nonetheless, the variation is mostly *within variants of EMAL* and is part of ablation study. We will separate ablation study from quantitative comparison in Table 1.

Q2: Details on how baselines perform adaptation.
A2: Domain adaptation methods minimize the discrepancy between the source and target features to adapt to the target domain. We modify DANN in EDA by sequentially adapting to target trajectories. JAN-merge and CDAN-merge merge the target trajectories as one target domain and adapt to it *offline*. We will provide more baseline implementation details.

Response to Reviewer 9

Q1: Motivation behind the combination of domain adaptation (DA) and online learning (OL) can be further elaborated.
A1: The proposed setting is *not a simple combination of DA and OL*. Evolving target data are ubiquitous in practice such as changing environment. *Forgetting is intrinsic* in this scenario. Saving target data may lead to *privacy issues*. Besides, evolving target data usually come *in small batches*, making our setting more challenging. We further provide a comparison of different settings of "continual" or "evolving" DA in Table 1.

Q2: The performance of the online setting is weak. The comparison against offline upper bound can be big.
A2: The performance in Table 1 of the main text seems weak since for each rotation we have *only 100 examples*, making it very challenging. **Offline** methods such as JAN-merge **cannot** be applied to the EDA setting directly, so we test it in offline setting. However, without capturing the information of evolvement, *adapting to target data offline as a whole hurts performance* (also observed in [14]). EAML even outperforms DANN+Replay (Replay Oracle in Table 1 of the main text) which involves heavy training of GANs for privacy-preserving Replay, justifying the efficacy and value of our method.

Q3: Comparison with MAML.
A3: We compare with *MAML adapted to EDA* in Table 3. Results indicate that the initialization learned with MAML does not provide *enough inductive bias* for EDA.