南京航空航天大學

# Decorate the Newcomers: Visual Domain Prompt for Continual Test Time Adaptation

**Yulu Gan[1], Yan Bai[1], Yihang Lou[2], Xianzheng Ma[3], Renrui Zhang[4], Nian Shi[5], Lin Luo[1]***

[1]Peking University, [2]Huawei Technologies, [3]Wuhan University, [4]The Chinese University of Hong Kong,
[5]Aerospace Information Research Institute, Chinese Academy of Sciences
ganyulu@stu.pku.edu.cn,{yanbai,zhangrenrui,luol}@pku.edu.cn,
louyihang1@huawei.com, maxianzheng@whu.edu.cn, shinian.work@gmail.com

AAAI 2023

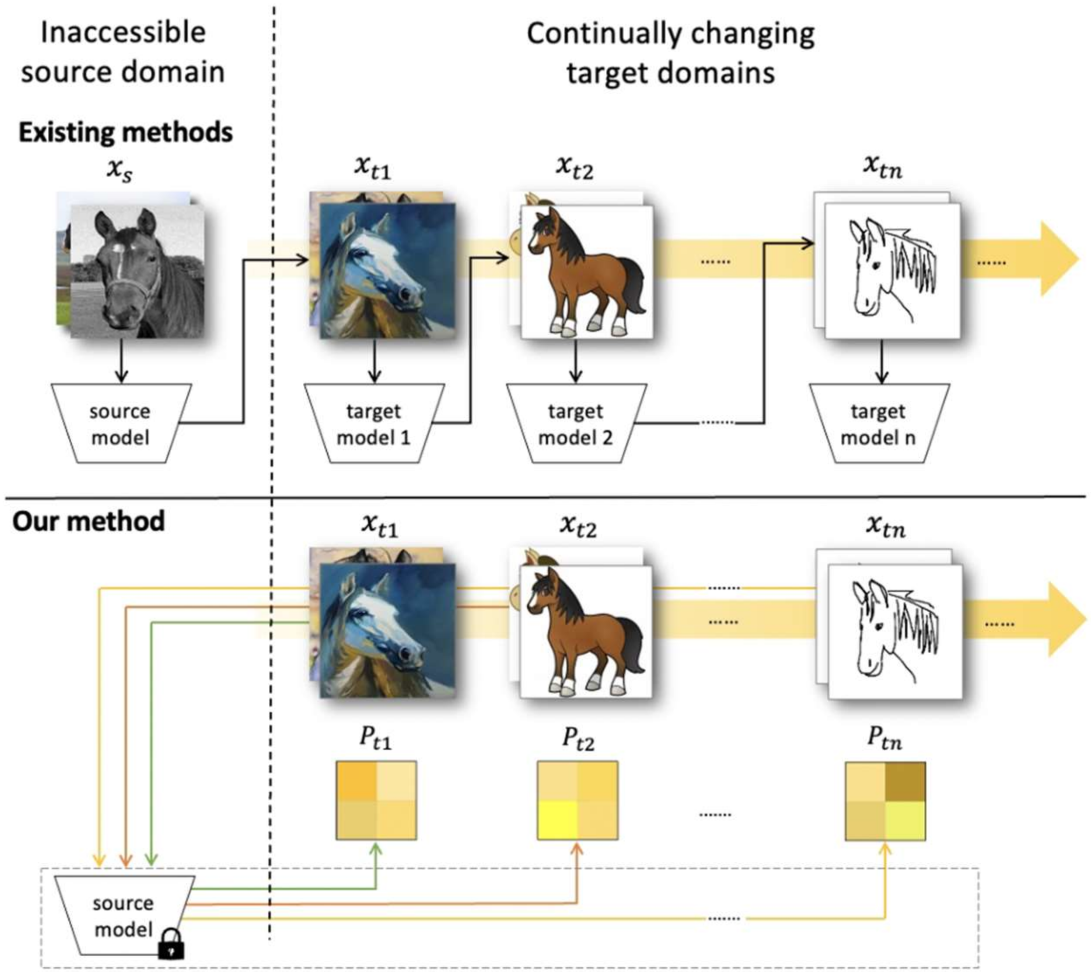
# Background

## □ Continual Test-Time Adaptation

- **Previous methods**

Focus on model-based adaptation

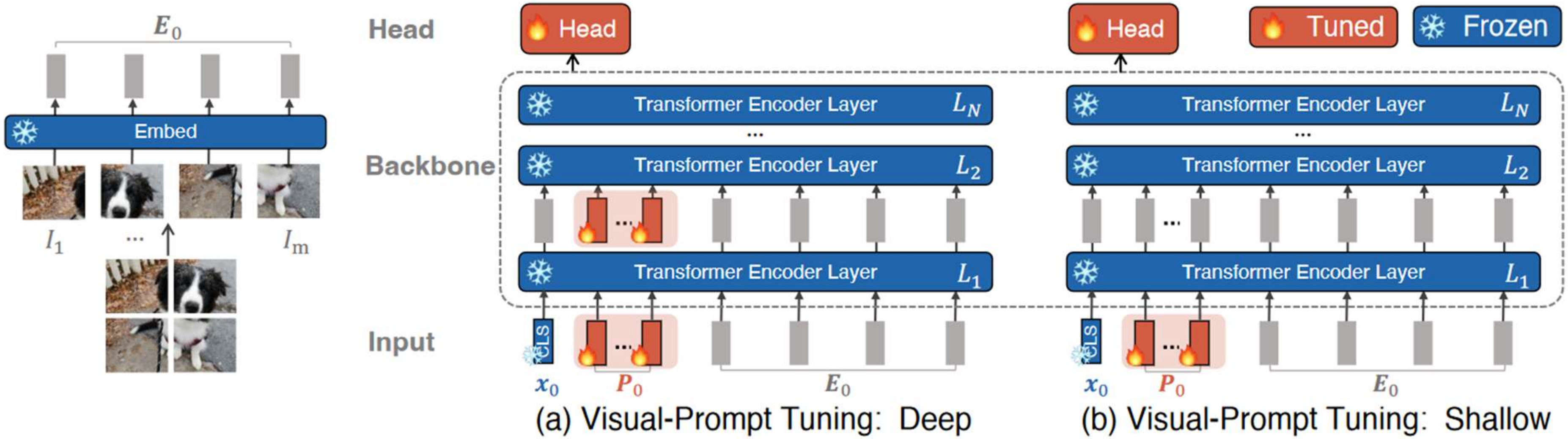
- **This work**
- Tuning the visual prompts for each domain
- Reformulating the input data with learned prompts



# Background

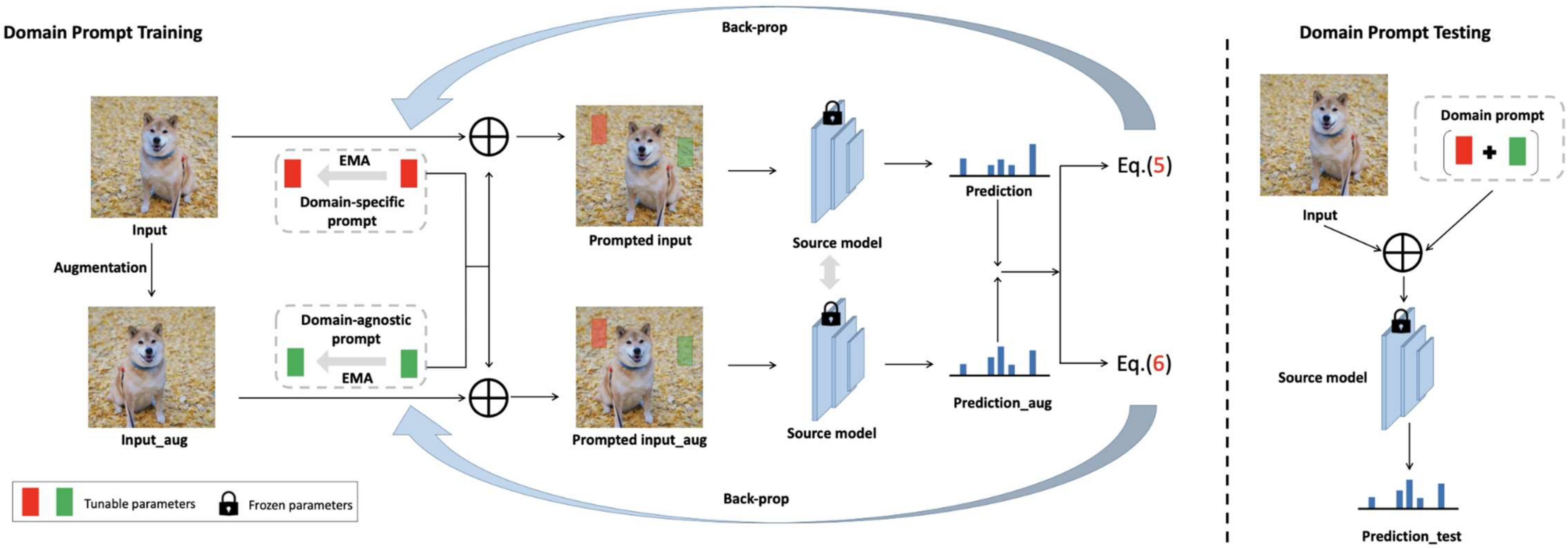


## ☐ Traditional visual prompts




Visual Prompt Tuning. ECCV 2022

## The whole framework




**Domain-specific prompts** → $$\mathcal{L}_{\omega_\phi}(x_p^T) = -\sum_C f_{\theta'_t}(h(x_p^T))(log f_{\theta t}(x_p^T)), \quad (5)$$

**Domain-agnostic prompts** → $$\mathcal{L}_{\psi_\delta}(x_p^T) = -\sum_C f_{\theta'_t}(h(x_p^T))(log f_{\theta t}(x_p^T)) + \boxed{\mathcal{L}(\psi_\delta),} \quad (6)$$

# Methods

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

## ☐ Limiting domain-sensitive parameters' update

$$\mathcal{L}(\psi_\delta) = \alpha \sum_{\theta \in \Theta} \Lambda_i^\tau ||\theta - \theta^*||_2^2,$$

$\theta^*$: model's parameters of the last mini-batch of the previous domain

→ **How to measure parameters' sensitivity toward domain shift?**

$t_0$ and $t_1$ respectively denote a certain time in two adjacent target domains

Inter-domain change
$$\begin{aligned}\mathcal{L}(\theta_{t_1}) - \mathcal{L}(\theta_{t_0}) &= \int_{t0}^{t1} g(\theta(t)) d\theta \\ &= \int_{t0}^{t1} g(\theta(t)) \cdot \theta'(t) dt \\ &= -\sum_i \eta_i^\nu,\end{aligned}$$

Intra-domain change
$$\delta_i^\nu = \theta_{i(t)}^\nu - \theta_{i(t-1)}^\nu$$

## ☐ Domain-shift detection

$$\Delta Conf = Conf_{t+1} - Conf_t$$

Threshold S=0.25

## ☐ The homeostatic factor

$$\Lambda_i^\tau = \sum_{v<\tau} \frac{\eta_i^\nu}{(\delta_i^\nu)^2 + \xi},$$

# Experiments

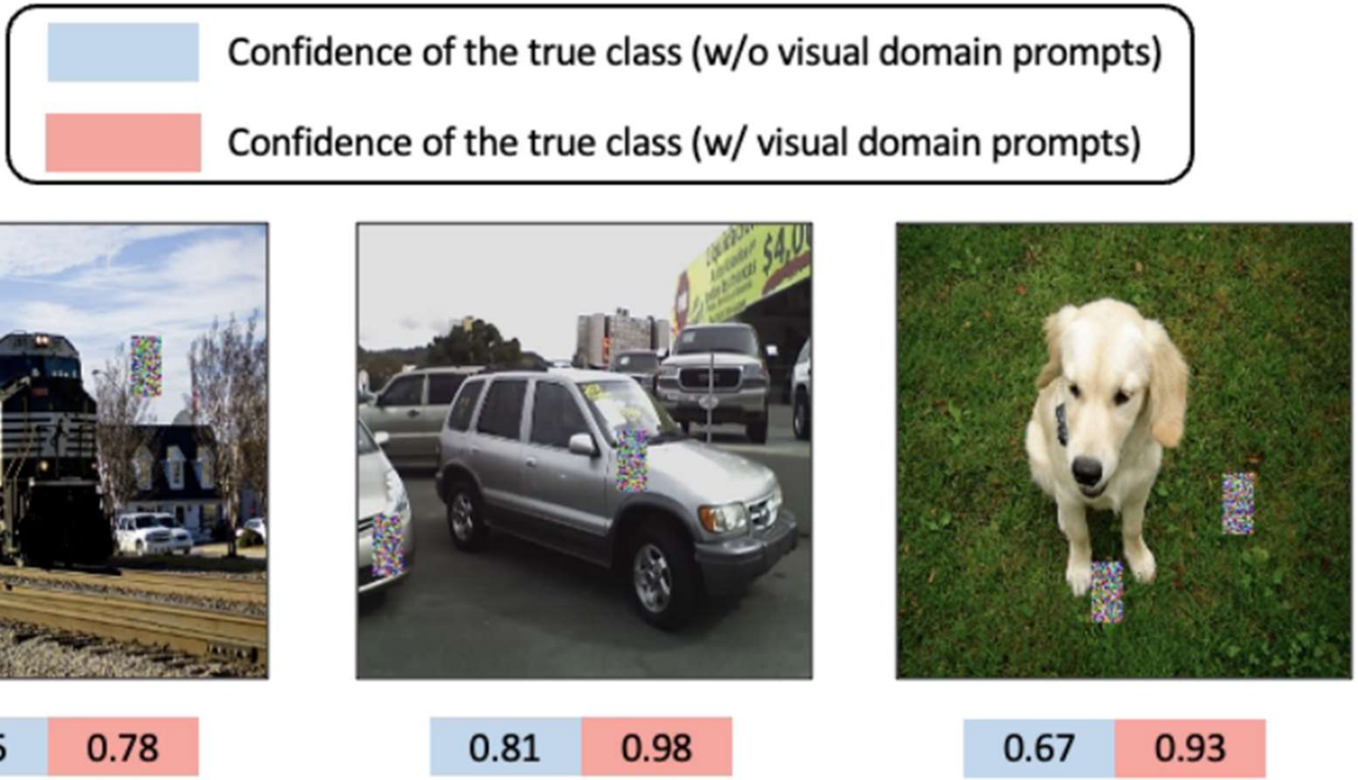
Table 2: Classification error rate (%) for the standard CIFAR100-to-CIFAR100C online continual test-time adaptation task. All results are evaluated on the ResNeXt-29 architecture with the largest corruption severity level 5. Our method far exceeds the state-of-the-art methods by 16.2%. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

| Method | gaussion | shot | impulse | defocus | glass | motion | zoom | snow | frost | fog | brightness | contrast | elastic_trans | pixelate | jpeg | Mean↓ | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 73.0 | 68.0 | 39.4 | 29.3 | 54.1 | 30.8 | 28.8 | 39.5 | 45.8 | 50.3 | 29.5 | 55.1 | 37.2 | 74.7 | 41.2 | 46.4 | 0.0 |
| BN Stats Adapt (Schneider et al. 2020) | 42.1 | 40.7 | 42.7 | 27.6 | 41.9 | 29.7 | 27.9 | 34.9 | 35.0 | 41.5 | 26.5 | 30.3 | 35.7 | 32.9 | 41.2 | 35.4 | +11.0 |
| Pseudo-label (Lee 2013) | 38.1 | 36.1 | 40.7 | 33.2 | 45.9 | 38.3 | 36.4 | 44.0 | 45.6 | 52.8 | 45.2 | 53.5 | 60.1 | 58.1 | 64.5 | 46.2 | +0.2 |
| Tent-continual (Wang et al. 2021a) | 37.2 | 35.8 | 41.7 | 37.9 | 51.2 | 48.3 | 48.5 | 58.4 | 63.7 | 71.1 | 70.4 | 82.3 | 88.0 | 88.5 | 90.4 | 60.9 | -14.5 |
| CoTTA (Wang et al. 2022a) | 40.1 | 37.7 | 39.7 | 26.9 | 38.0 | 27.9 | 26.4 | 32.8 | 31.8 | 40.3 | 24.7 | 26.9 | 32.5 | 28.3 | 33.5 | 32.5 | +13.9 |
| **Ours (proposed)** | **29.5** | **25.8** | **31.9** | **2.8** | **30.5** | **7.7** | **5.7** | **14.8** | **14.8** | **24.2** | **1.8** | **6.8** | **18.5** | **9.1** | **28.0** | **16.8** | **+29.6** |

Table 3: Classification error rate(%) for standard CIFAR10-to-CIAFAR10C online continual test-time adaptation task. Results are evaluated on WideResNet-28 with the largest corruption severity level 5. Our method exceeds the state-of-the-art methods by 2.3%. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

| Method | gaussion | shot | impulse | defocus | glass | motion | zoom | snow | frost | fog | brightness | contrast | elastic_trans | pixelate | jpeg | Mean↓ | Gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 72.3 | 65.7 | 72.9 | 46.9 | 54.3 | 34.8 | 42.0 | 25.1 | 41.3 | 26.0 | 9.3 | 46.7 | 26.6 | 58.5 | 30.3 | 43.5 | 0.0 |
| BN Stats Adapt (Schneider et al. 2020) | 28.1 | 26.1 | 36.3 | 12.8 | 35.3 | 14.2 | 12.1 | 17.3 | 17.4 | 15.3 | 8.4 | 12.6 | 23.8 | 19.7 | 27.3 | 20.4 | +23.1 |
| Pseudo-label (Lee 2013) | 26.7 | 22.1 | 32.0 | 13.8 | 32.2 | 15.3 | 12.7 | 17.3 | 17.3 | 16.5 | 10.1 | 13.4 | 22.4 | 18.9 | 25.9 | 19.8 | +23.7 |
| Tent-online (Wang et al. 2021a) | 24.8 | 23.5 | 33.0 | 12.0 | 31.8 | 13.7 | 10.8 | 15.9 | 16.2 | 13.7 | 7.9 | 12.1 | 22.0 | 17.3 | 24.2 | 18.6 | +24.9 |
| Tent-continual (Wang et al. 2021a) | 24.8 | 20.6 | 28.6 | 14.4 | 31.1 | 16.5 | 14.1 | 19.1 | 18.6 | 18.6 | 12.2 | 20.3 | 25.7 | 20.8 | 24.9 | 20.7 | +22.8 |
| Baseline(CoTTA) (Wang et al. 2022a) | 24.3 | 21.3 | **26.6** | 11.6 | **27.6** | 12.2 | 10.3 | 14.8 | 14.1 | 12.4 | 7.5 | 10.6 | **18.3** | 13.4 | **17.3** | 16.2 | +27.3 |
| **Ours (proposed)** | **22.6** | **19.7** | 28.1 | **7.1** | 28.4 | **9.5** | **6.3** | **10.2** | **11.5** | **9.0** | **1.5** | **5.6** | 18.5 | **12.8** | 18.5 | **13.9** | **+29.6** |

# Experiments



# Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 7: **Ablation: Contribution of our proposed DAP and DSP.**

| # | DSP | DAP | CIFAR10C | CIFAR100C | ImageNet-C |
|---|---|---|---|---|---|
| 0 | | | 16.2 | 32.5 | 63.0 |
| 1 | ✓ | | 14.9 | 17.8 | 52.4 |
| 2 | | ✓ | 15.2 | 19.5 | 53.2 |
| 3 | ✓ | ✓ | 13.9 | 16.3 | 51.5 |

$$\mathcal{L}_{\omega_\phi}(x_p^T) = -\sum_C f_{\theta'_t}(h(x_p^T))(log f_{\theta_t}(x_p^T)),$$

$$\mathcal{L}_{\psi_\delta}(x_p^T) = -\sum_C f_{\theta'_t}(h(x_p^T))(log f_{\theta_t}(x_p^T)) + \boxed{\mathcal{L}(\psi_\delta),}$$

# Experiments

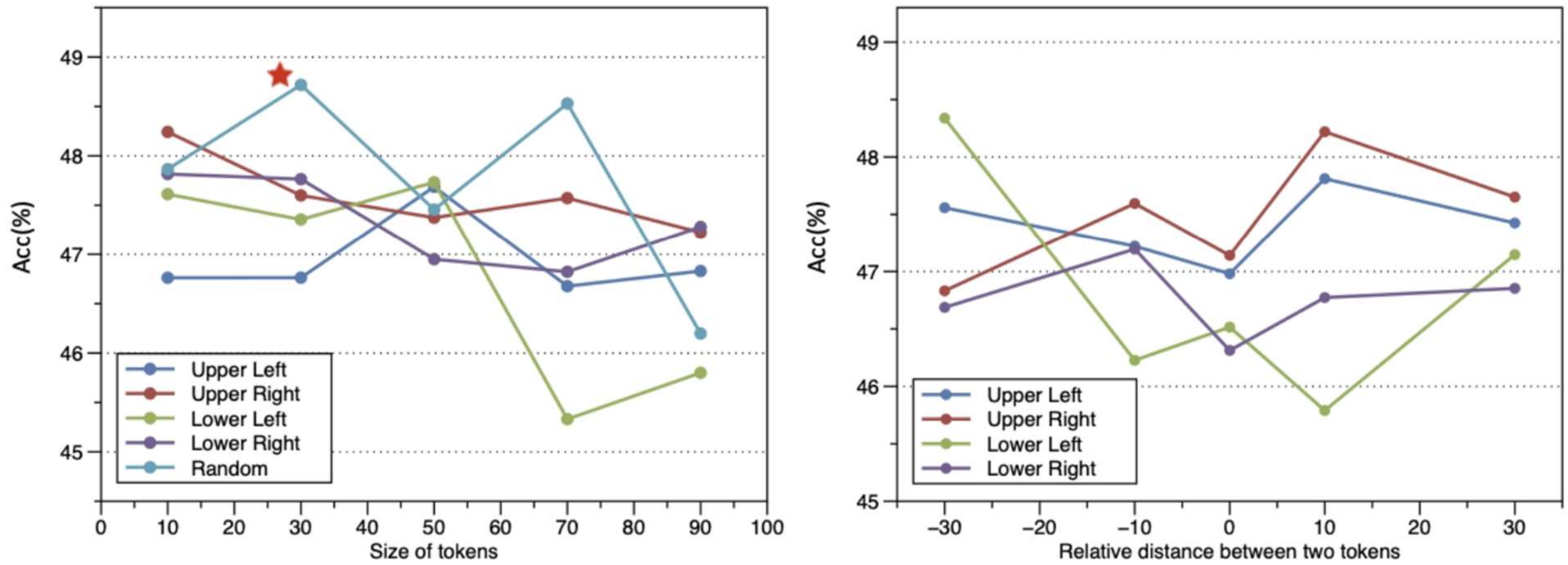



Figure 4: **Effects of the prompts' size, location on the image and relative distance between the two prompts.** Experiments are conducted on the ImageNet-to-ImageNet-C task. The left figure shows the effect of the size and position of prompts on the model performance. When the prompts' size equals 30, and apply to the image randomly, the model's performance achieves the best. The right figure shows the effect of the relative positions. We set the prompt size to $20\times 20$. If the distance is negative, the domain-independent prompt will be on the left side of the domain-specific prompt and vice versa. We find that exchanging the positions of the two prompts and altering the distance between these two prompts will affect the model's performance.

# Dataset Condensation with Gradient Matching

**Bo Zhao, Konda Reddy Mopuri, Hakan Bilen**
School of Informatics, The University of Edinburgh
{bo.zhao, kmopuri, hbilen}@ed.ac.uk

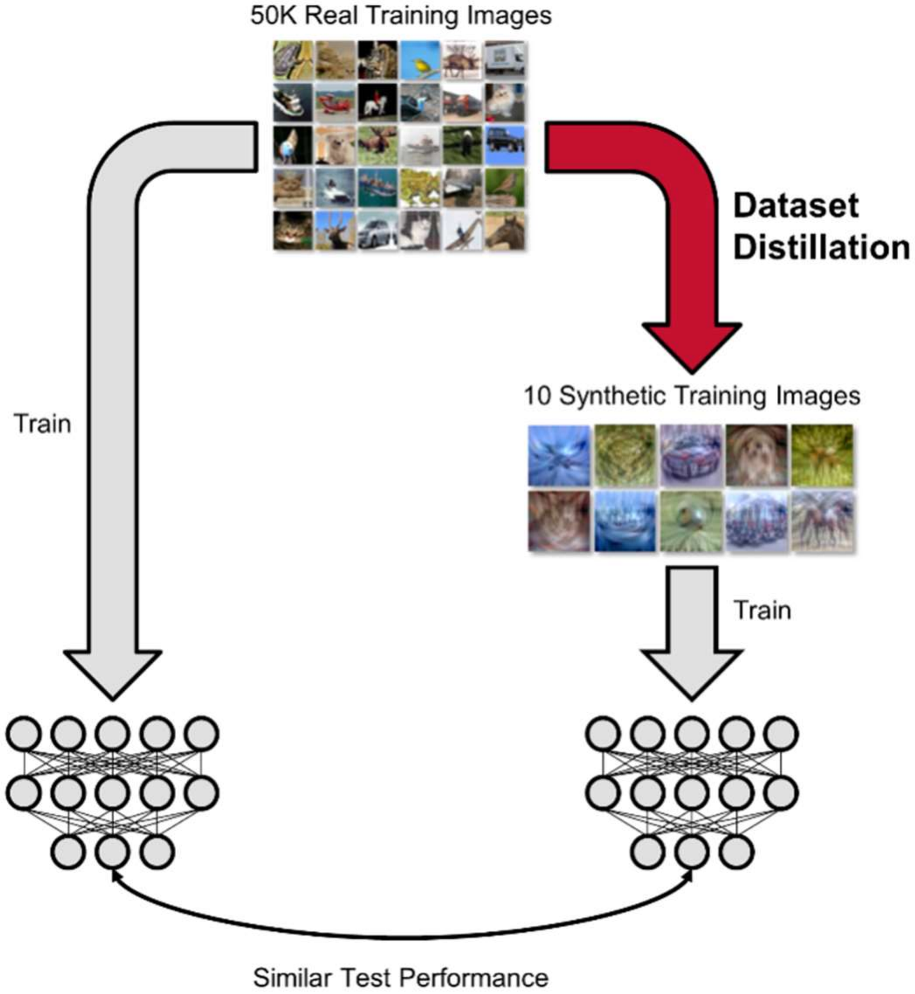
ICLR 2021

# Background

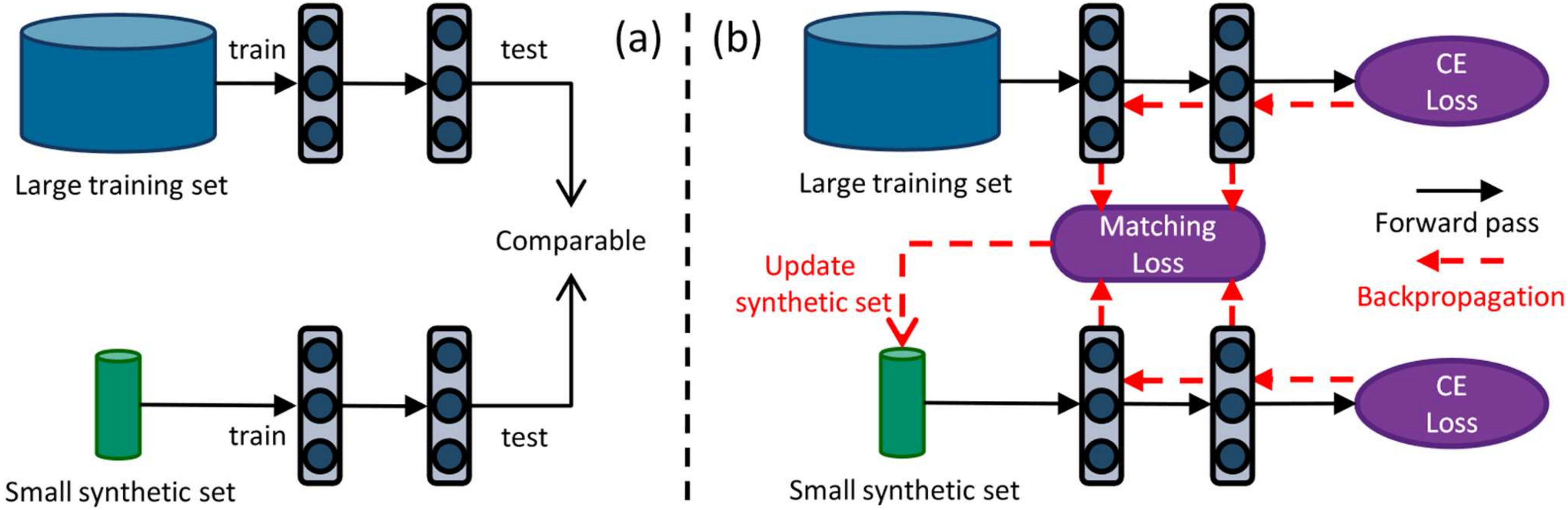


## Dataset Distillation

**Task:** Synthesizing small datasets such that models trained on them achieve comparable performance to the original large dataset.



## □ The whole framework



# Methods

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

**Goal:** Synthetic data $S^*$

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{L}^{\mathcal{T}}(\boldsymbol{\theta}^{\mathcal{S}}(\mathcal{S})) \qquad \text{subject to} \qquad \boldsymbol{\theta}^{\mathcal{S}}(\mathcal{S}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}).$$

## Parameter Matching

- Similar weights imply similar mappings in a local neighborhood and thus generalization performance

$$\min_{\mathcal{S}} D(\boldsymbol{\theta}^{\mathcal{S}}, \boldsymbol{\theta}^{\mathcal{T}}) \quad \text{subject to} \quad \boldsymbol{\theta}^{\mathcal{S}}(\mathcal{S}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta})$$

- To generate samples that can work with a distribution of random initializations

$$\min_{\mathcal{S}} \mathrm{E}_{\boldsymbol{\theta}_0 \sim P_{\boldsymbol{\theta}_0}}[D(\boldsymbol{\theta}^{\mathcal{S}}(\boldsymbol{\theta}_0), \boldsymbol{\theta}^{\mathcal{T}}(\boldsymbol{\theta}_0))] \quad \text{subject to} \quad \boldsymbol{\theta}^{\mathcal{S}}(\mathcal{S}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}(\boldsymbol{\theta}_0))$$

Re-defines $\theta^S$ as the output of an incomplete optimization $\boldsymbol{\theta}^{\mathcal{S}}(\mathcal{S}) = \texttt{opt-alg}_{\boldsymbol{\theta}}(\mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}), \varsigma)$

Justin Domke. Generic methods for optimization-based modeling. In Artificial Intelligence and Statistics. 2012.

# Methods

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

## Curriculum Gradient Matching

- Let $\theta^S$ follow a similar path to $\theta^T$ throughout the optimization

$$\min_{\mathcal{S}} \mathrm{E}_{\boldsymbol{\theta}_0 \sim P_{\boldsymbol{\theta}_0}} [\sum_{t=0}^{T-1} D(\boldsymbol{\theta}_t^{\mathcal{S}}, \boldsymbol{\theta}_t^{\mathcal{T}})] \quad \text{subject to}$$

$$\boldsymbol{\theta}_{t+1}^{\mathcal{S}}(\mathcal{S}) = \texttt{opt-alg}_{\boldsymbol{\theta}}(\mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}_t^{\mathcal{S}}), \varsigma^{\mathcal{S}}) \quad \text{and} \quad \boldsymbol{\theta}_{t+1}^{\mathcal{T}} = \texttt{opt-alg}_{\boldsymbol{\theta}}(\mathcal{L}^{\mathcal{T}}(\boldsymbol{\theta}_t^{\mathcal{T}}), \varsigma^{\mathcal{T}})$$

- Minimize the distance between the gradients

$$\boldsymbol{\theta}_{t+1}^{\mathcal{S}} \leftarrow \boldsymbol{\theta}_t^{\mathcal{S}} - \eta_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}_t^{\mathcal{S}}) \quad \text{and} \quad \boldsymbol{\theta}_{t+1}^{\mathcal{T}} \leftarrow \boldsymbol{\theta}_t^{\mathcal{T}} - \eta_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{T}}(\boldsymbol{\theta}_t^{\mathcal{T}}),$$

$$\min_{\mathcal{S}} \mathrm{E}_{\boldsymbol{\theta}_0 \sim P_{\boldsymbol{\theta}_0}} [\sum_{t=0}^{T-1} D(\nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}(\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{T}}(\boldsymbol{\theta}_t))].$$

$$D(\nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{S}}, \nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathcal{T}}) = \textstyle\sum_{l=1}^{L} d(\nabla_{\boldsymbol{\theta}^{(l)}} \mathcal{L}^{\mathcal{S}}, \nabla_{\boldsymbol{\theta}^{(l)}} \mathcal{L}^{\mathcal{T}}) \qquad d(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{\text{out}} \left(1 - \frac{\mathbf{A_{i \cdot}} \cdot \mathbf{B_{i \cdot}}}{\|\mathbf{A_{i \cdot}}\| \|\mathbf{B_{i \cdot}}\|}\right)$$

# Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

| | Img/Cls | Ratio % | Coreset Selection Random | Herding | K-Center | Forgetting | Ours | Whole Dataset |
|---|---|---|---|---|---|---|---|---|
| MNIST | 1 | 0.017 | 64.9±3.5 | 89.2±1.6 | 89.3±1.5 | 35.5±5.6 | **91.7±0.5** | 99.6±0.0 |
| | 10 | 0.17 | 95.1±0.9 | 93.7±0.3 | 84.4±1.7 | 68.1±3.3 | **97.4±0.2** | |
| | 50 | 0.83 | 97.9±0.2 | 94.9±0.2 | 97.4±0.3 | 88.2±1.2 | **98.8±0.2** | |
| FashionMNIST | 1 | 0.017 | 51.4±3.8 | 67.0±1.9 | 66.9±1.8 | 42.0±5.5 | **70.5±0.6** | 93.5±0.1 |
| | 10 | 0.17 | 73.8±0.7 | 71.1±0.7 | 54.7±1.5 | 53.9±2.0 | **82.3±0.4** | |
| | 50 | 0.83 | 82.5±0.7 | 71.9±0.8 | 68.3±0.8 | 55.0±1.1 | **83.6±0.4** | |
| SVHN | 1 | 0.014 | 14.6±1.6 | 20.9±1.3 | 21.0±1.5 | 12.1±1.7 | **31.2±1.4** | 95.4±0.1 |
| | 10 | 0.14 | 35.1±4.1 | 50.5±3.3 | 14.0±1.3 | 16.8±1.2 | **76.1±0.6** | |
| | 50 | 0.7 | 70.9±0.9 | 72.6±0.8 | 20.1±1.4 | 27.2±1.5 | **82.3±0.3** | |
| CIFAR10 | 1 | 0.02 | 14.4±2.0 | 21.5±1.2 | 21.5±1.3 | 13.5±1.2 | **28.3±0.5** | 84.8±0.1 |
| | 10 | 0.2 | 26.0±1.2 | 31.6±0.7 | 14.7±0.9 | 23.3±1.0 | **44.9±0.5** | |
| | 50 | 1 | 43.4±1.0 | 40.4±0.6 | 27.0±1.4 | 23.3±1.1 | **53.9±0.5** | |

# Experiments

| C\T | MLP | ConvNet | LeNet | AlexNet | VGG | ResNet |
|---|---|---|---|---|---|---|
| MLP | 70.5±1.2 | 63.9±6.5 | 77.3±5.8 | 70.9±11.6 | 53.2±7.0 | 80.9±3.6 |
| ConvNet | 69.6±1.6 | **91.7±0.5** | 85.3±1.8 | 85.1±3.0 | **83.4±1.8** | **90.0±0.8** |
| LeNet | 71.0±1.6 | 90.3±1.2 | 85.0±1.7 | 84.7±2.4 | 80.3±2.7 | 89.0±0.8 |
| AlexNet | 72.1±1.7 | 87.5±1.6 | 84.0±2.8 | 82.7±2.9 | 81.2±3.0 | 88.9±1.1 |
| VGG | 70.3±1.6 | 90.1±0.7 | 83.9±2.7 | 83.4±3.7 | 81.7±2.6 | 89.1±0.9 |
| ResNet | **73.6±1.2** | 91.6±0.5 | **86.4±1.5** | **85.4±1.9** | **83.4±2.4** | 89.4±0.9 |

Table 2: Cross-architecture performance in testing accuracy (%) for condensed 1 image/class in MNIST.

# Thanks