Semi-Supervised Fine-Tuning of Vision Foundation Models with Content-Style Decomposition

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITÉ DE GENÈVE

Mariia Drozdova, Vitaliy Kinakh, Yury Belousov, Erica Lastufka, Slava Voloshynovskiy

Université de Genève

{mariia.drozdova, vitaliy.kinakh, yury.belousov, erica.lastufka, svolos}@unige.ch

Introduction

Foundation models, pre-trained on vast datasets, show strong generalization but struggle with:

- Distribution shifts: Downstream data often differs from pre-training data.
- Low-labeled regimes: Labeled data is costly in domains like astronomy and medicine.

Our approach uses **semi-supervised fine-tuning** with a novel **content-style decomposition** framework to enhance adaptation to downstream tasks with limited labeled data.

Motivation

Scientific domains often have:

- Abundant unlabeled data but limited labeled samples.
- Tasks requiring adaptation to domain-specific datasets (e.g., GalaxyMNIST, SVHN).

Key Question: Can unlabeled data help fine-tune foundation models for distribution-shifted downstream tasks?

Datasets

MNIST

MNIST with yellow strips

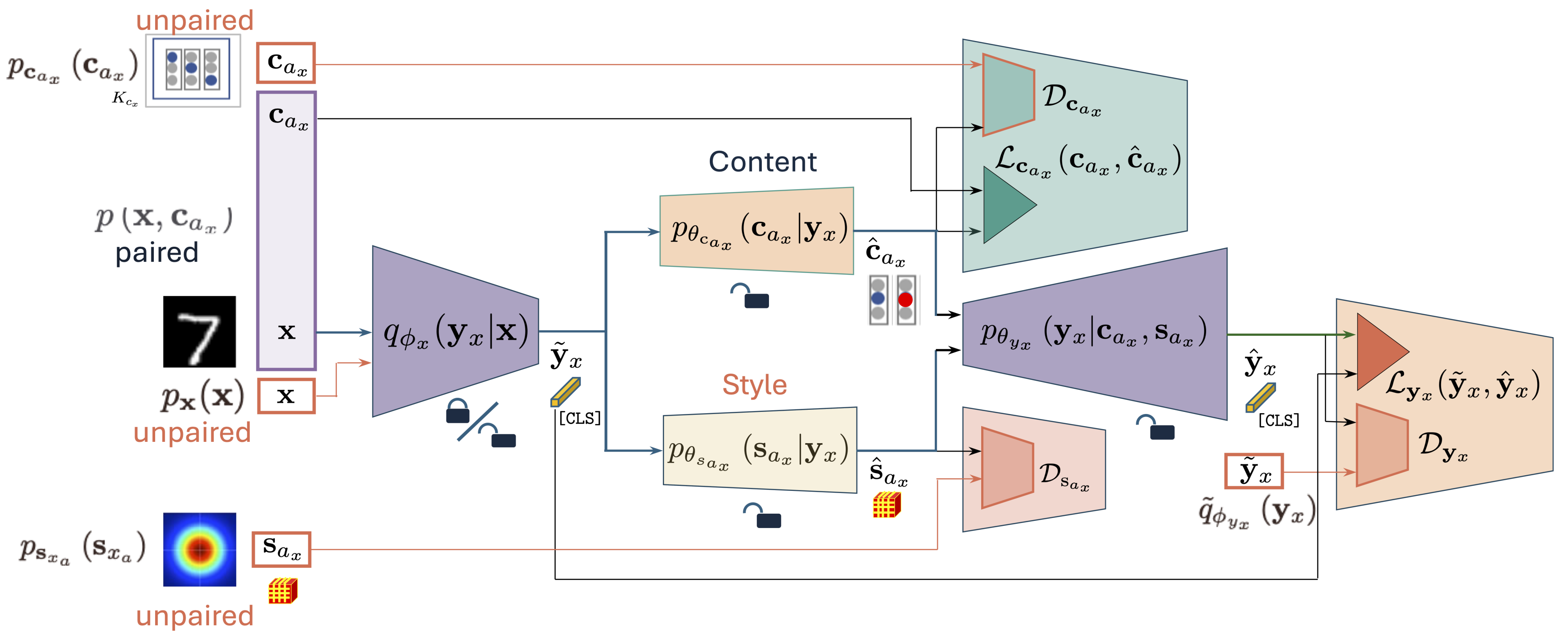
MNIST with white strips

SVHN

CIFAR-10

GalaxyMNIST

Approach


unpaired
$p_{\mathbf{c}_{a_x}}(\mathbf{c}_{a_x})$
K_{c_x}
$\mathbf{c}_{a_x}$
$\mathbf{c}_{a_x}$
$p(\mathbf{x}, \mathbf{c}_{a_x})$
paired
$\mathbf{x}$
$p_{\mathbf{x}}(\mathbf{x})$
$\mathbf{x}$
unpaired
$q_{\phi_x}(\mathbf{y}_x|\mathbf{x})$
$\tilde{\mathbf{y}}_x$
[CLS]
Content
$p_{\theta_{c_{a_x}}}(\mathbf{c}_{a_x}|\mathbf{y}_x)$
$\hat{\mathbf{c}}_{a_x}$
Style
$p_{\theta_{s_{a_x}}}(\mathbf{s}_{a_x}|\mathbf{y}_x)$
$\hat{\mathbf{s}}_{a_x}$
$\mathcal{D}_{\mathbf{c}_{a_x}}$
$\mathcal{L}_{\mathbf{c}_{a_x}}(\mathbf{c}_{a_x}, \hat{\mathbf{c}}_{a_x})$
$p_{\theta_{y_x}}(\mathbf{y}_x|\mathbf{c}_{a_x}, \mathbf{s}_{a_x})$
$\mathcal{D}_{\mathbf{s}_{a_x}}$
$\hat{\mathbf{y}}_x$
[CLS]
$\tilde{\mathbf{y}}_x$
$\tilde{q}_{\phi_{y_x}}(\mathbf{y}_x)$
$\mathcal{L}_{\mathbf{y}_x}(\tilde{\mathbf{y}}_x, \hat{\mathbf{y}}_x)$
$\mathcal{D}_{\mathbf{y}_x}$
$p_{\mathbf{s}_{x_a}}(\mathbf{s}_{x_a})$
$\mathbf{s}_{a_x}$
unpaired

The [CLS] token $\mathbf{y}_x$ is decomposed into:

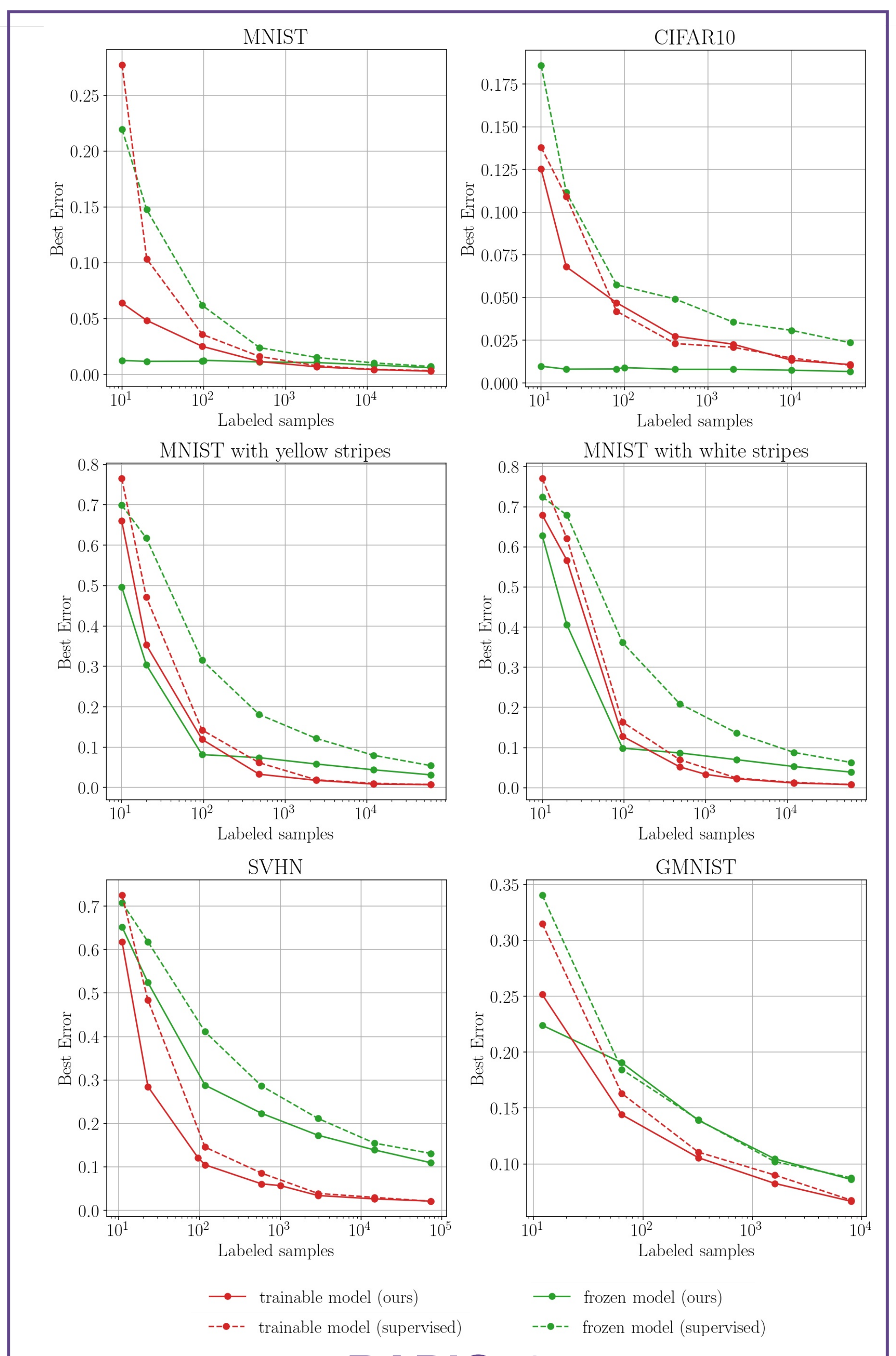
- Content $\mathbf{c}_{a_x}$: Task-specific class information (one-hot encoding).
- Style $\mathbf{s}_{a_x}$: Auxiliary Gaussian noise representing variations.

Key Losses:

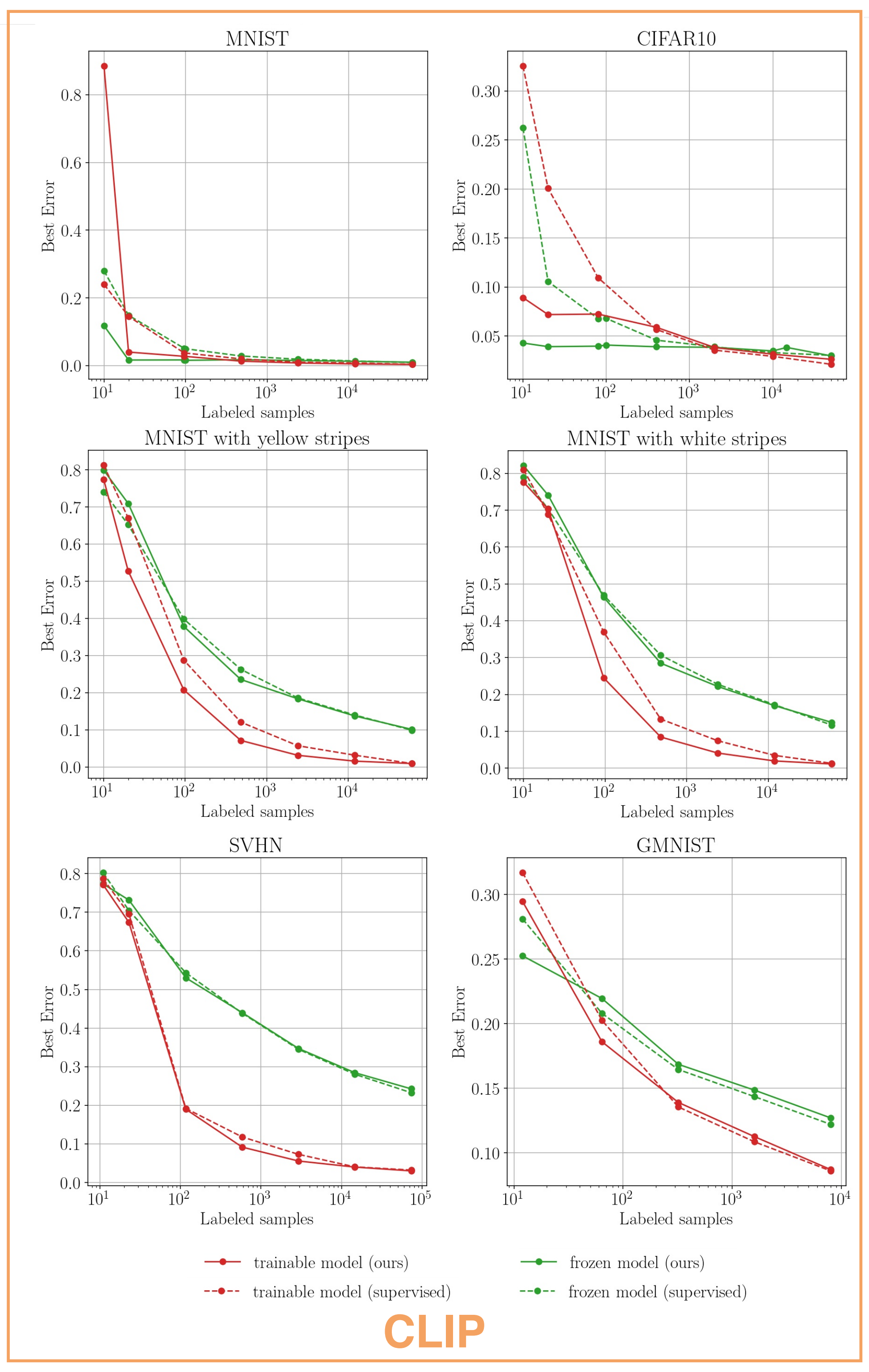
- Cross-Entropy for content prediction: $\mathcal{L}_{\mathbf{c}_{a_x}}$.
- Cosine Similarity for CLS reconstruction: $\mathcal{L}_{\mathbf{y}_x}$.
- KL Divergences for regularizing content, style and reconstruction distributions: $\mathcal{D}_{\mathbf{c}_{a_x}}$, $\mathcal{D}_{\mathbf{s}_{a_x}}$, $\mathcal{D}_{\mathbf{y}_x}$.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathbf{c}_{a_x}} + \lambda_c \mathcal{D}_{\mathbf{c}_{a_x}} + \lambda_s \mathcal{D}_{\mathbf{s}_{a_x}} + \lambda_y \mathcal{D}_{\mathbf{y}_x} + \lambda_{y\hat{y}} \mathcal{L}_{\mathbf{y}_x}$$

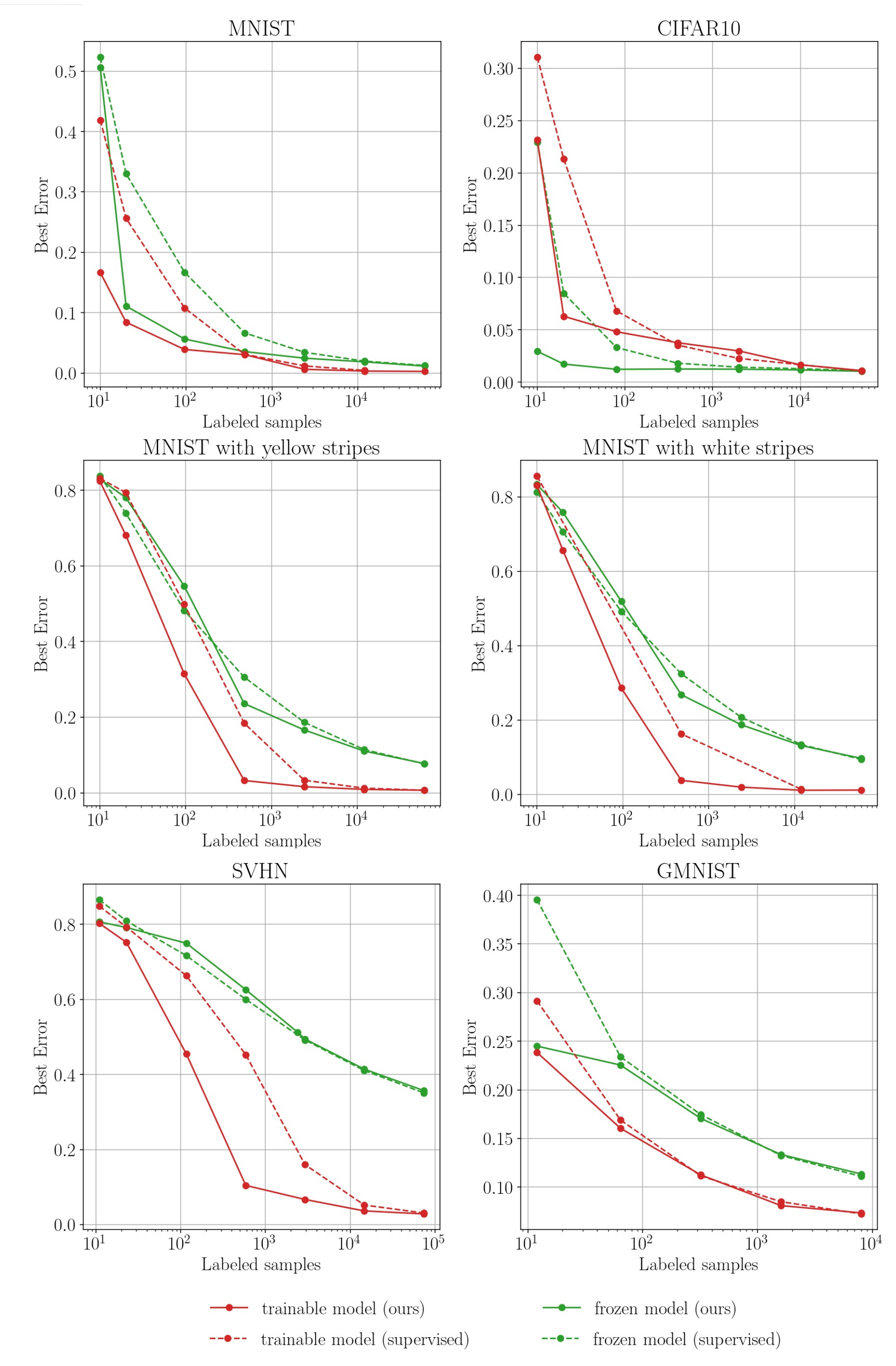
Experiments



RADIOv2



CLIP



DINOv2

Models:
- RADIOv2
- CLIP
- DINOv2

Scenarios:
- Frozen backbones (only classifier is fine-tuned)
- Trainable backbones (full model updates)

Evaluation: Compare supervised and semi-supervised fine-tuning.

Regularizing the latent space (content and style) ensures task-relevant information is preserved while discarding noise.

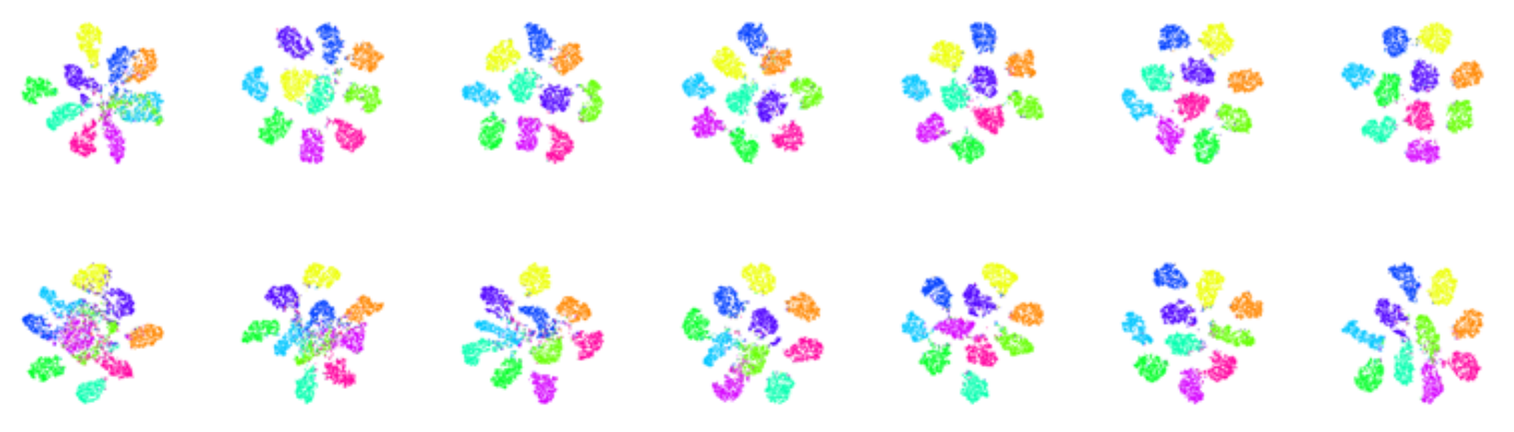
Frozen vs. Trainable:
Frozen backbones excel in low-labeled regimes.Trainable models adapt better for complex tasks with more labeled data.

Conclusions

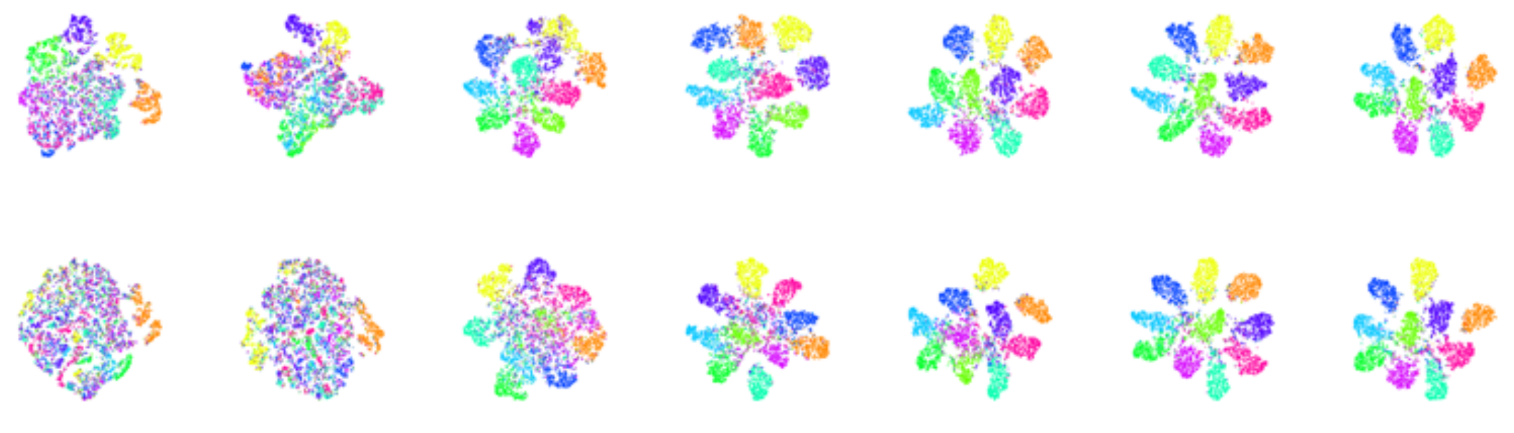
- It improves performance in **low-labeled regimes** by leveraging unlabeled data effectively..
- Semi-supervised fine-tuning bridges the gap between pre-trained models and downstream tasks, especially for **out-of-distribution data**.
- **In the Future** we would like to test classification on larger and more complex datasets like DomainNet and ImageNet variations and to apply the method to object detection and segmentation tasks.

TSNEs

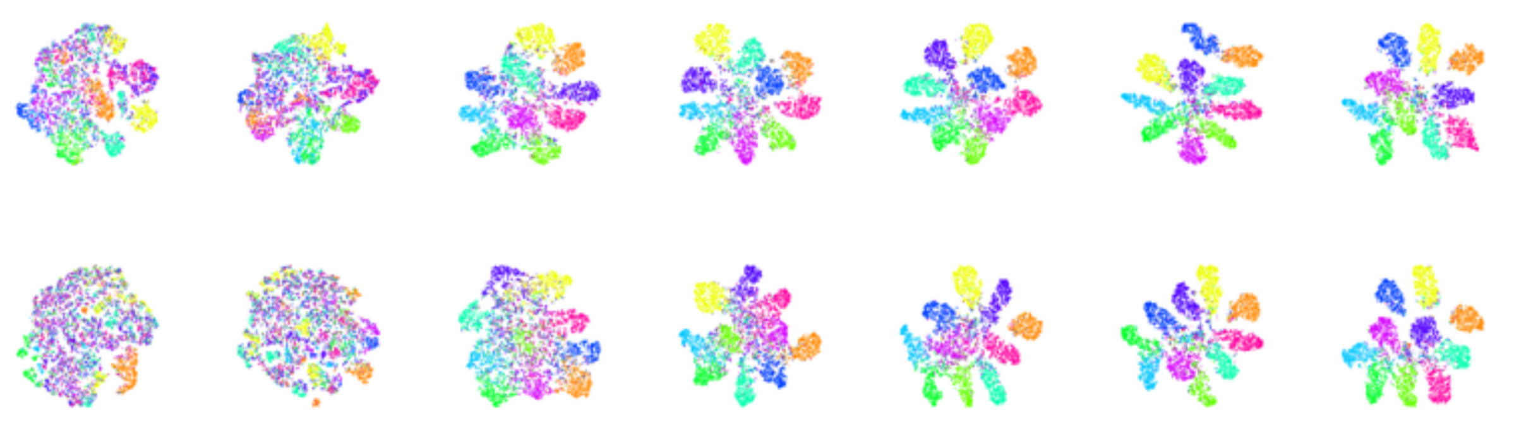
MNIST



MNIST with yellow strips



MNIST with white strips



References

Slava Voloshynovskiy, Olga Taran, Mouad Kondah, Taras Holotyak, and Danilo Rezende.Variational information bottleneck for semi-supervised classification. In Entropy, 22(9):943, 2020

arXiv:2410.02069