NEURAL INFORMATION PROCESSING SYSTEMS

Semi-Supervised Contrastive Learning for Deep Regression with Ordinal Rankings from Spectral Seriation

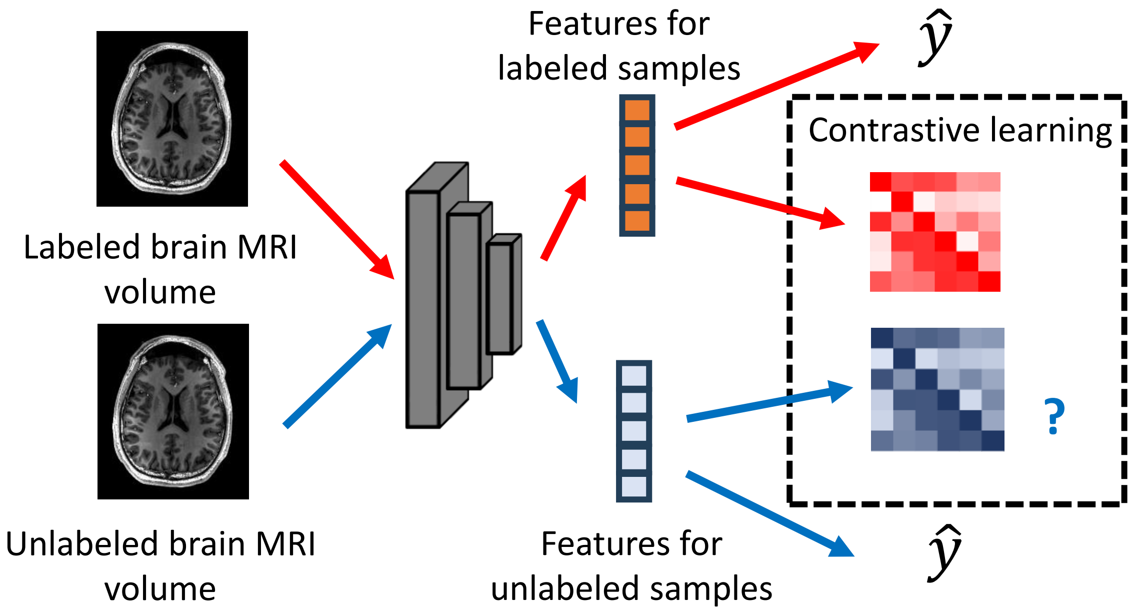
NeurIPS 2023

Weihang Dai[1], Yao Du[1], Hanru Bai[3], Kwang-Ting Cheng[1], Xiaomeng Li[1,2]*

[1]The Hong Kong University of Science and Technology
[2]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen
[3]Fudan University

Background

- Contrastive learning is a state-of-the-art feature learning technique applied to classification and regression
- Can be used on unlabeled data for classification tasks and for unsupervised model pre-training
- CanNOT be used on unlabeled data for regression tasks

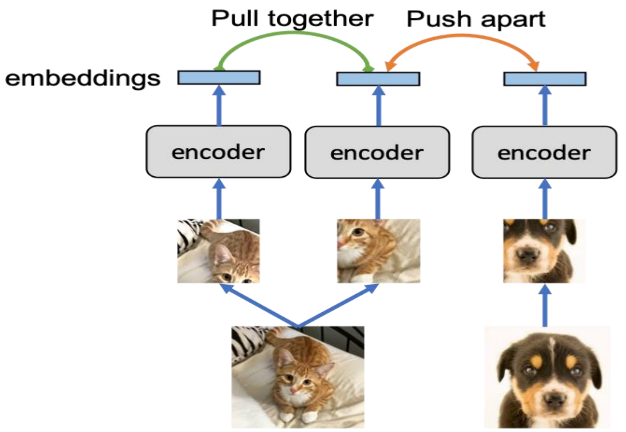


Can we extend contrastive learning methods for regression to a semi-supervised setting?

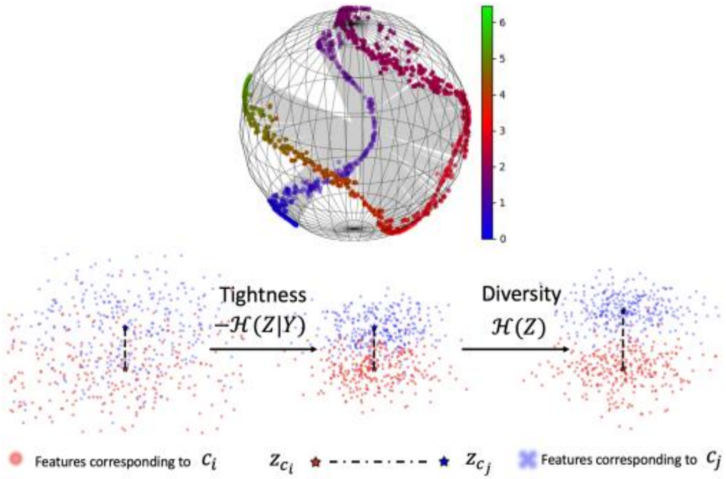
Related Works

- Contrastive learning methods for classification
  - SimCLR, MOCO
  - Set positive pairs as augmented inputs of same sample
- Contrastive learning methods for regression
  - AdaCon, Ordinal Entropy [1]
  - Requires labels in order to enforce distance relationships

Unsupervised Contrastive Learning



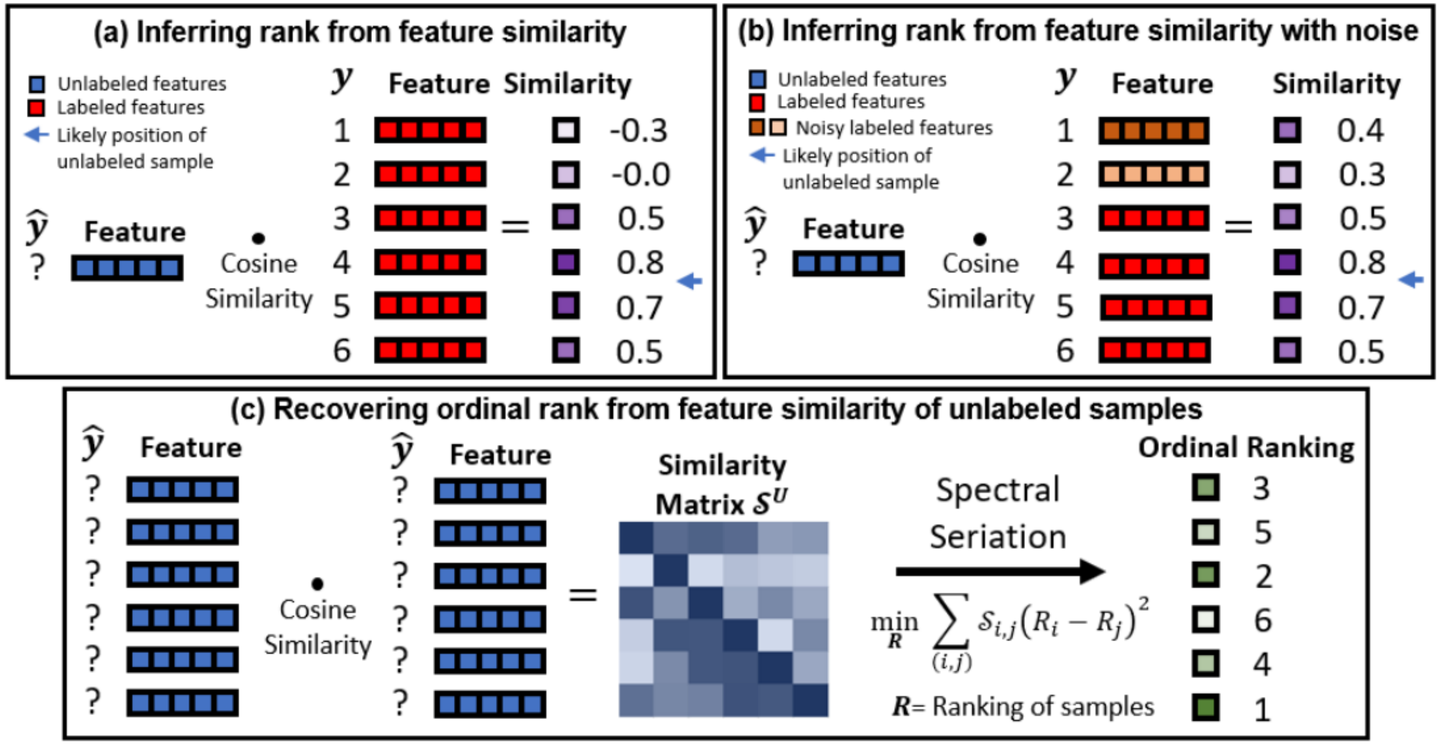
Ordinal Entropy



[1] Zhang, Shihao, et al. "Improving Deep Regression with Ordinal Entropy." *arXiv preprint arXiv:2301.08915* (2023).

Main Idea – Recovering rankings from similarity matrix

- Contrastive learning leads features on unlabeled samples to also reflect label distance
- We can recover the ranking of unlabeled samples from their noisy similarity matrix through spectral seriation



Method – The Spectral Seriation Algorithm

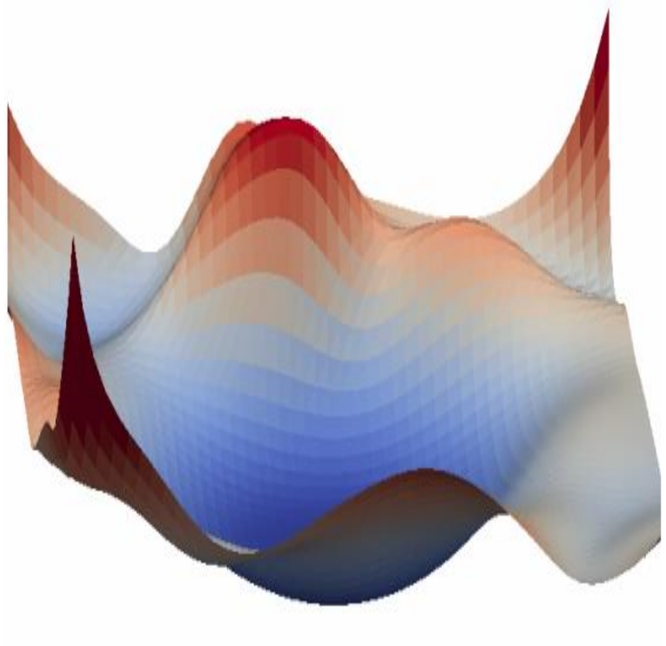
Finding optimal ranking given similarity matrix

- For similarity matrix S, where samples closer together have higher similarity values, spectral seriation finds the most likely ranking of samples
- Spectral seriation minimizes the following:

$$\underset{R}{\operatorname{argmin}} \sum_{i,j} S_{i,j} \left(R_i - R_j\right)^2 ,$$

  where R is the rankings, $S_{i,j}$ are entries in similarity matrix S

- R is obtained through loss minimization and can therefore be robust to noise



The optimum point can be robust to shifts and perturbations in the surface

Theorem 2 *For a similarity matrix $S' \in \mathbb{R}^{n\times n}$, suppose the error matrix of it is $E \in \mathbb{R}^{n\times n}$. When*

$$||E||_1 \leq \frac{\lambda_3 - \lambda_2}{8\sqrt{n}},$$

where λ_2, λ_3 are the second smallest and the third smallest eigenvalue of Laplacian matrix of S', the Fiedler vector of $S' \in \mathbb{R}^{n\times n}$ is stable, so the seriation obtained by the spectral ranking algorithm is robust to noise in S'.

Method – Constraining feature and predictions for unlabeled samples

Constraining similarity matrix and predictions for unlabeled samples

- **Constraining similarity matrix for unlabeled samples**
  - We can use recovered sample rankings to constrain our similarity matrix for contrastive learning
  - We ensure similarity values with respect to sample i follow the same ordering inferred from derived rankings

$$\mathcal{L}^{UC} = \sum_{i=1}^{|\mathcal{B}|} \ell\left(\mathbf{rk}(\mathcal{S}'_{[i,:]}),\ \mathbf{rk}(-|R' - R'_{[i]}|); \lambda\right),$$

where *[i, :]* denotes the ith row in the matrix, *[i]* denotes the *i*th value of a vector, **rk** denotes the ranking operator, and ℓ is the ranking similarity function.

Method – Constraining feature and predictions for unlabeled samples

Constraining similarity matrix and predictions for unlabeled samples

- **Constraining similarity matrix for unlabeled samples**
  - We can use recovered sample rankings to constrain our similarity matrix for contrastive learning
  - We ensure similarity values with respect to sample i follow the same ordering inferred from derived rankings

$$\mathcal{L}^{UC} = \sum_{i=1}^{|\mathcal{B}|} \ell \left(\mathbf{rk}(\mathcal{S}'_{[i,:]}), \mathbf{rk}(-|R' - R'_{[i]}|); \lambda \right),$$

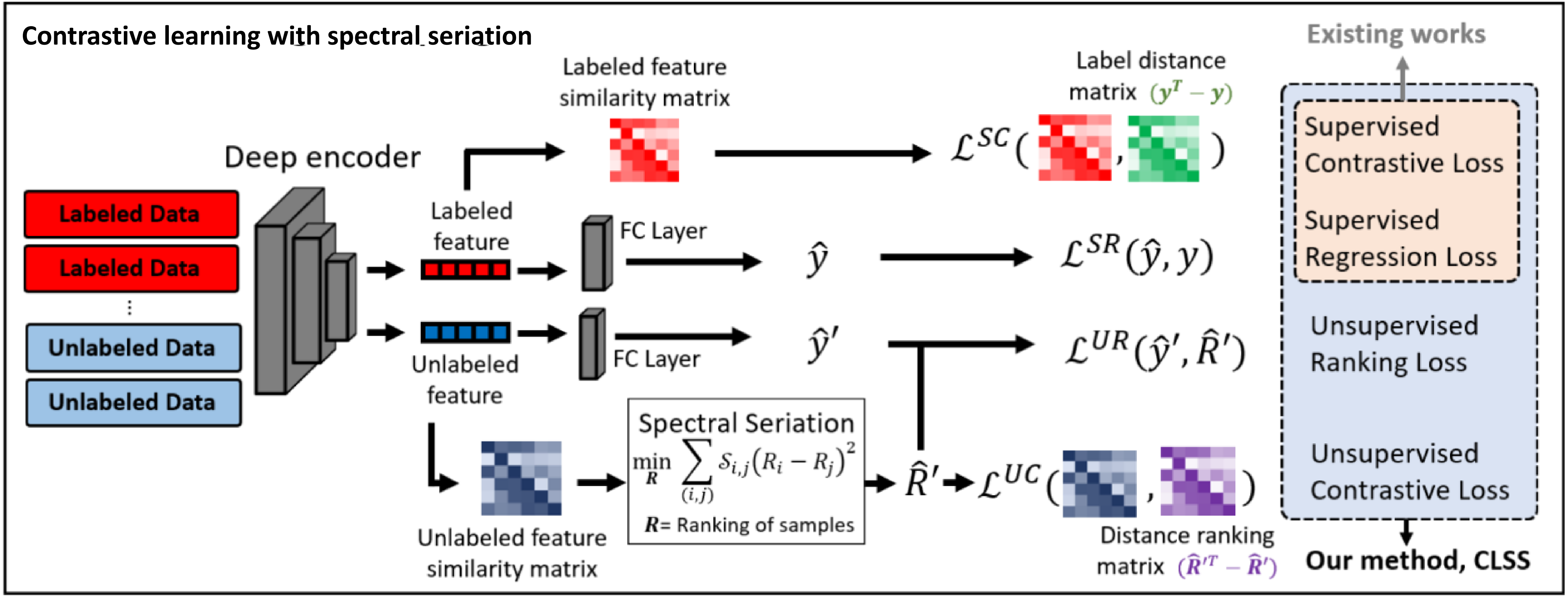
where *[i, :]* denotes the ith row in the matrix, *[i]* denotes the ith value of a vector, **rk** denotes the ranking operator, and ℓ is the ranking similarity function

- **Constraining predictions for unlabeled samples**
  - Rankings from spectral seriation are error tolerant and can also be used to supervise predictions

$$\mathcal{L}^{UR} = \sum_{i=1}^{|\mathcal{B}|} \ell \left(\mathbf{rk}(-|\hat{y}' - \hat{y}'_{[i]}|), \mathbf{rk}(-|R' - R'_{[i]}|); \lambda \right),$$

where *[i]* denotes the *i*th value of a vector

Method – Overall Framework (CLSS)

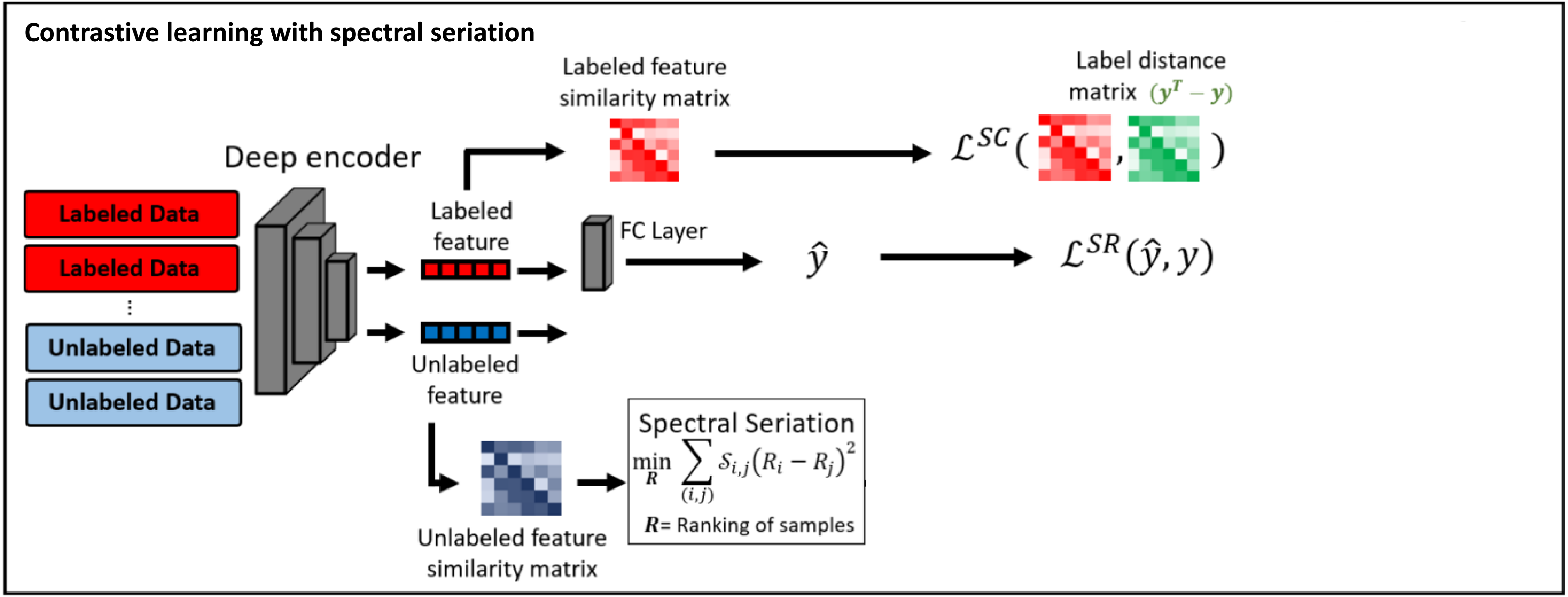


The total loss function $\mathcal{L}$ of our method is:

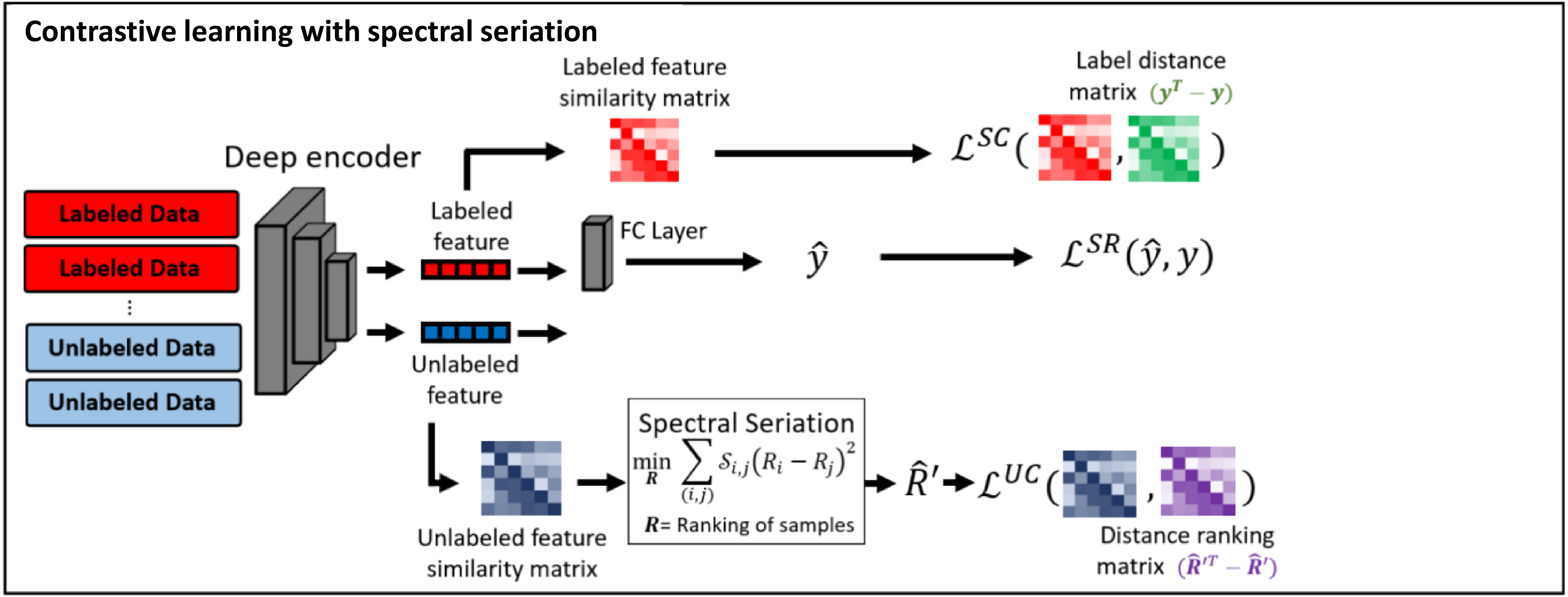
$$\mathcal{L} = \mathcal{L}^{SR} + w_{SC}\mathcal{L}^{SC} + w_{UC}\mathcal{L}^{UC} + w_{UR}\mathcal{L}^{UR},$$

where $\mathcal{L}^{SR}$, $\mathcal{L}^{SC}$, $\mathcal{L}^{UC}$ and $\mathcal{L}^{UR}$ represent the loss values of supervised regression, supervised contrastive loss, unsupervised contrastive loss, and unsupervised ranking loss. w_{SC}, w_{UC} and w_{UR} are the corresponding loss weights

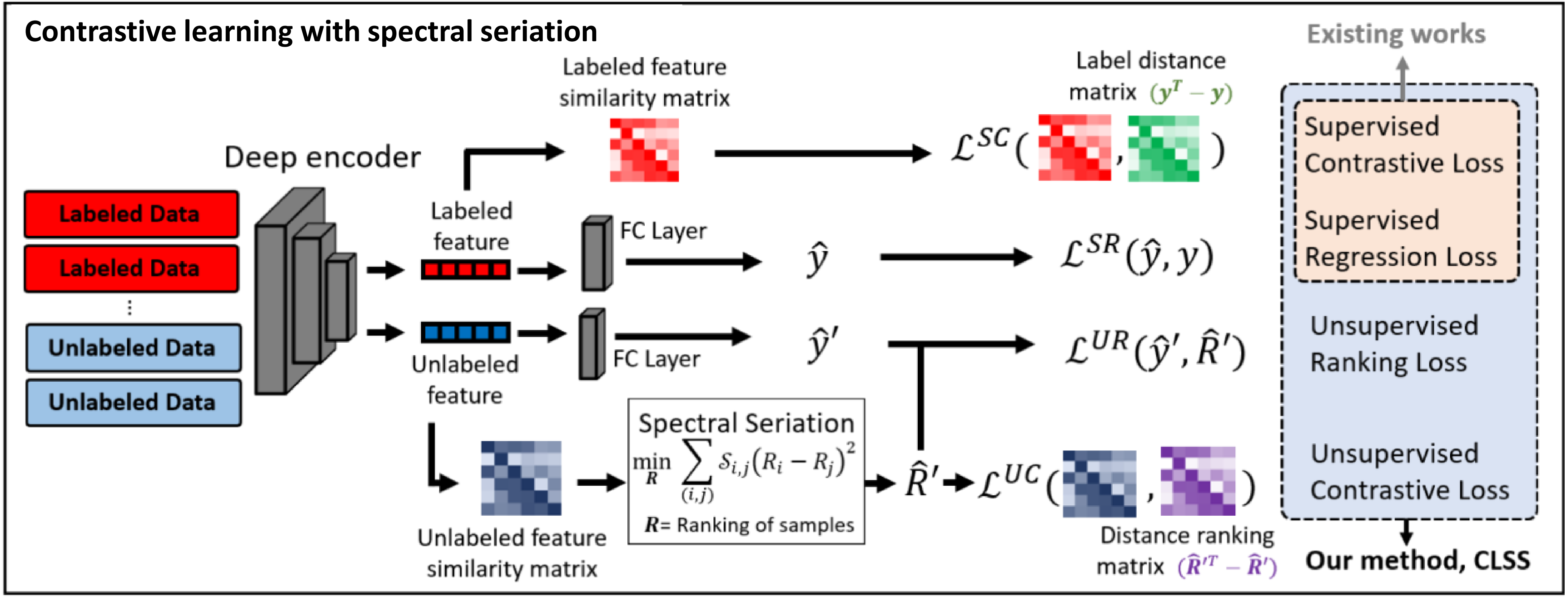
Method – Overall Framework (CLSS)



Method – Overall Framework (CLSS)



Method – Overall Framework (CLSS)



Results – Experiments

- Validation on Brain Age estimation from MRI Scans

Comparison with state-of-the-art on IXI brain age estimation dataset

Type	Method	MAE↓ 1/5 labels	1/4 labels	1/3 labels	1/2 labels
Supervised	Regression	9.95 ± 1.41	11.93 ± 1.40	11.76 ± 1.75	10.93 ± 1.60
Semi-supervised	Mean-teacher	11.23 ± 2.31	10.27 ± 1.57	10.52 ± 3.12	12.01 ± 2.03
	CPS	10.23 ± 1.41	10.27 ± 1.19	**9.64 ± 1.27**	9.69 ± 1.01
	UCVME	9.83 ± 1.32	10.86 ± 1.67	9.65 ± 1.31	10.06 ± 1.19
	CLSS (Ours)	**9.58 ± 1.48**	**9.68 ± 1.22**	9.72 ± 1.29	**9.37 ± 1.17**

CLSS leads to more stable results and reduces reliance on healthy patients for labeled data

Results – Experiments

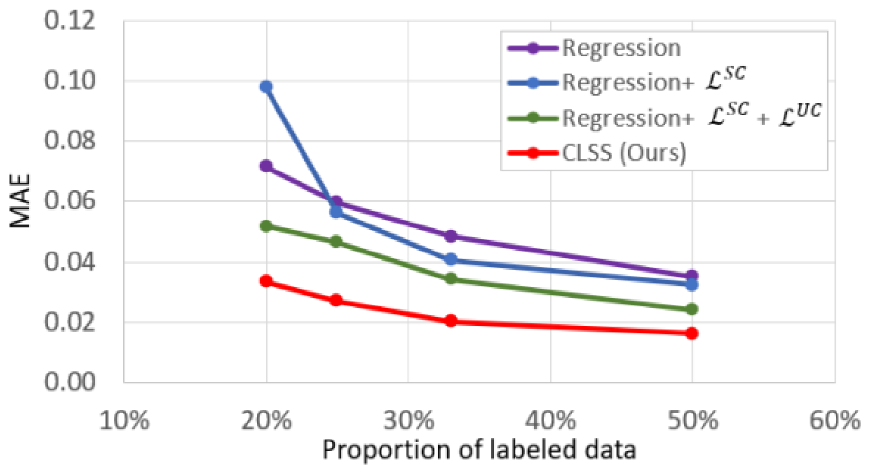
- Synthetic dataset for non-linear operator learning

We train a model to solve the following PDE:

$$-\mathrm{div}(e^{b(x;w)}\nabla u(x;w)) = f(x)$$

Comparison with state-of-the-art on synthetic PDE dataset

		MAE↓			
Type	Method	1/5 labels	1/4 labels	1/3 labels	1/2 labels
Supervised	Regression	0.098 ± 0.095	0.056 ± 0.016	0.041 ± 0.015	0.032 ± 0.009
Semi-supervised	Mean-teacher	0.080 ± 0.089	0.047 ± 0.021	0.043 ± 0.019	0.029 ± 0.011
	CPS	0.057 ± 0.012	0.045 ± 0.016	0.041 ± 0.015	0.028 ± 0.007
	UCVME	0.040 ± 0.008	0.033 ± 0.008	0.027 ± 0.007	0.028 ± 0.021
	CLSS (Ours)	**0.033 ± 0.008**	**0.027 ± 0.009**	**0.020 ± 0.007**	**0.016 ± 0.007**



CLSS outperforms state-of-the-art semi-supervised deep regression methods for all settings

Results – Experiments

- Validation on Age-Estimation from photographs

Comparison with state-of-the-art methods on AgeDB-DIR dataset

		MAE↓			
Type	Method	1/30 labels	1/25 labels	1/20 labels	1/15 labels
Supervised	Regression	10.14 ± 0.25	9.99 ± 0.11	9.10 ± 0.15	8.58 ± 0.10
Semi-supervised	Mean-teacher [28]	10.05 ± 0.29	9.99 ± 0.13	9.05 ± 0.12	8.62 ± 0.09
	CPS [4]	9.99 ± 0.12	9.83 ± 0.10	8.99 ± 0.14	8.47 ± 0.08
	Ours	$\mathbf{9.95 \pm 0.18}$	$\mathbf{9.59 \pm 0.12}$	$\mathbf{8.88 \pm 0.09}$	$\mathbf{8.45 \pm 0.11}$

- CLSS outperforms state-of-the-art semi-supervised deep regression methods for all settings
- CLSS can also be applied effectively to natural image datasets