Korea university Biomedical Engineering

SNUH RADIOLOGY

CCIDS 의료영상데이터사이언스센터 Center for Clinical Imaging Data Science

2020. 04. 12

Unsupervised domain adaptation for medical imaging segmentation with self-ensembling

Sang Wook Kim
Korea university
Department of Biomedical Engineering

Christian S. Perone., “Unsupervised domain adaptation for medical imaging segmentation with self-ensembling”, 2018

Contents

Introduction

1. The large amount of required data to train DNN => Partially mitigated with techniques such as transfer learning
    1. However, transfer learning is problematic in medical imaging **b.c a large dataset is still required**
    2. Medical images require careful and **time-consuming analysis from trained experts**
2. One of the most difficult to solve is the **so-called data distribution shift**
    1. Models trained under the **empirical risk minimization** (ERM) principle, might **fail to generalize** to other domains **due to its strong assumptions**
    2. ERM is good if its assumptions hold the fact **such as the fact that the training and test datasets derive from similar domains**
3. Distribution Shift is a problem in medical image analysis
    1. Over-optimistic evaluation results

Introduction

3. **Distribution Shift** is a problem in medical image analysis
    1. Over-optimistic evaluation results
4. The name given to learn a classifier model with a shift btw the training and the target distribution is called as "**domain adaptation**"
5. **Contributions**
    1. Extend **unsupervised DA** method using **self-ensembling**
    2. Ablation experiment to provide **strong evidence that unlabeled data is responsible** for the observed performance improvement
    3. Provide **visualizations** using t-SNE

Introduction

[*] **Unsupervised domain adaptation** (UDA) is the task of **training a statistical model on labeled data from a source domain to achieve better performance on data from a target domain, with access to only unlabeled data in the target domain**

reference : https://www.aclweb.org/anthology/N19-1039/

Related Work

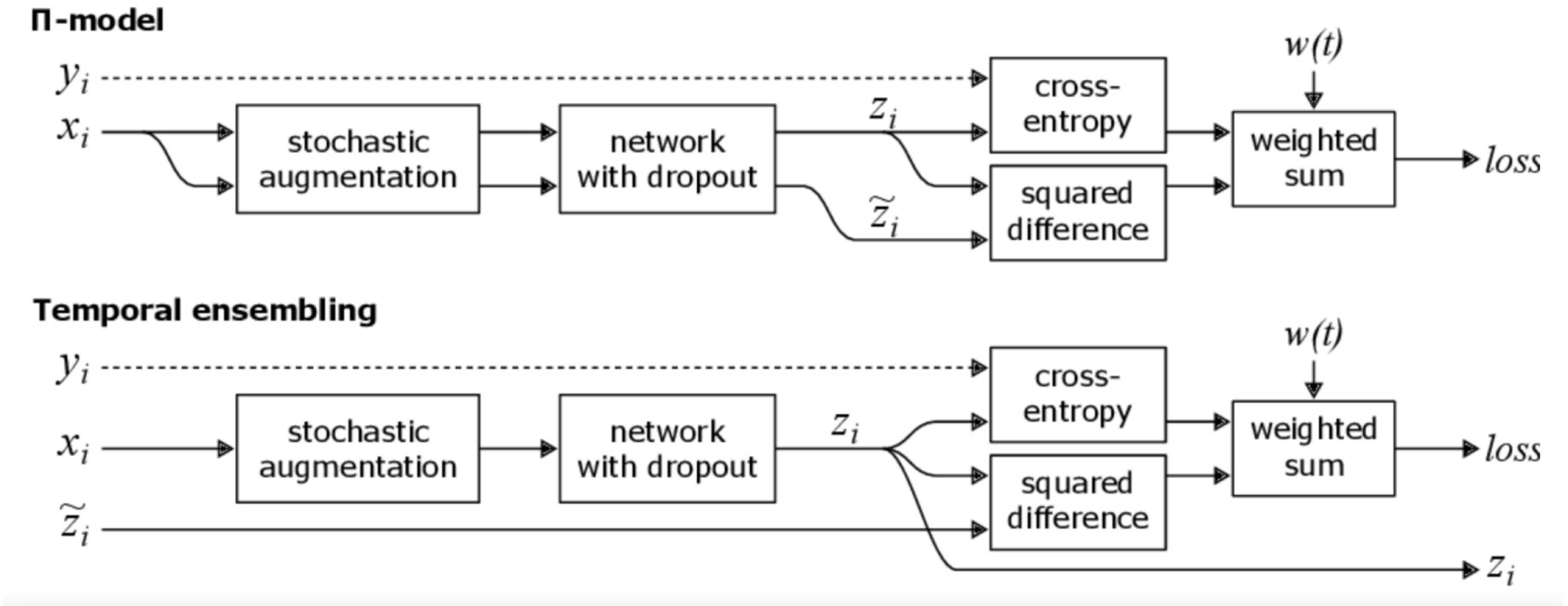
1. **Deep Domain Adaptation (DDA)**
    1. Literatures
        1. Auto-encoders, GANs, Disentanglement strategies
        2. Higher-order statistics
        3. Explicit discrepancy btw source and target domain
        4. **Self-ensembling methods based on implicit discrepancy**
            1. Widely used for unsupervised domain adaptation
            2. Based on the **Mean Teacher network** (= first introduced for semi-supervised learning task)
            3. Very few adjustments that need to be made to employ the method for the purpose of DDA

Semi-supervised learning and unsupervised domain adaptation

1. Semi-supervised learning
    1. The assumption that the distribution of unlabeled data is similar to labeled data often fails in real-world applications
    2. **Domain shift** : the **difference** btw the distributions from the examples used in **training and test set**
2. **Unsupervised Domain adaptation**
    1. Only P(Xs), P(Y|Xs), P(Xt) is available (No P(Y|Xt))
    2. The task is **to leverage knowledge from the target domain using the unlabeled data available in P(Xt)**

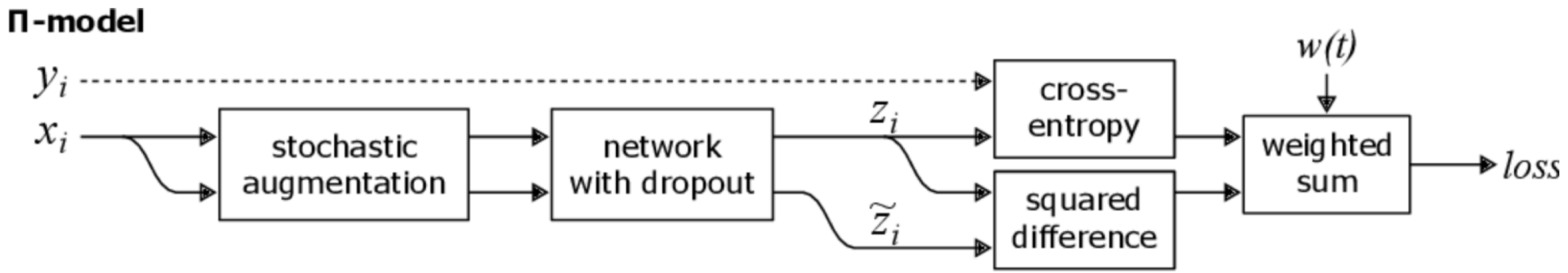
Self-ensembling and mean teacher

1. Self-ensembling



Related Work

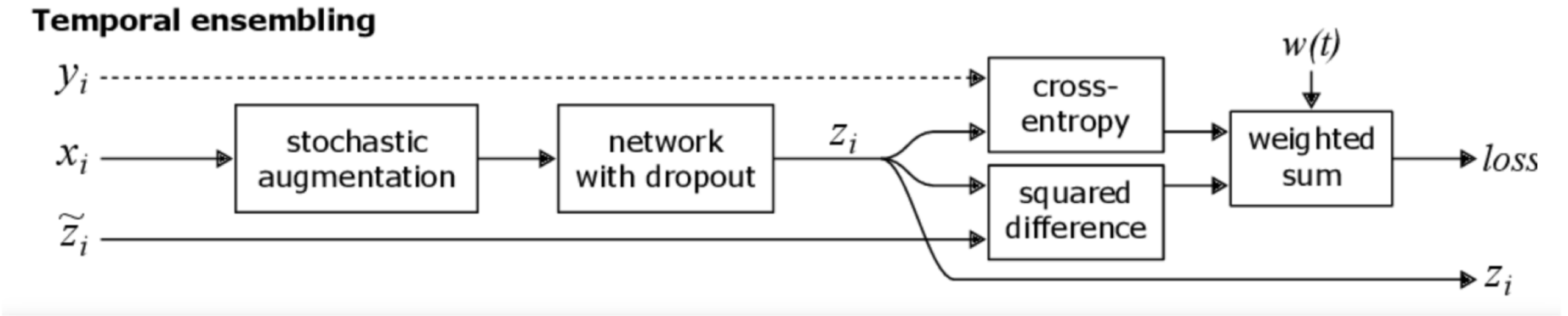
1. Π (pi)
    1. Employs a **consistency loss** btw predictions on the same input
        1. Each input from a batch is **passed twice through a NN**
        2. Each time with **distinct augmentation** parameters **to yield two diff. Predictions**
        3. **A squared diff.** btw those predictions is minimized along with the cross-entropy for the labeled examples



Related Work

2. Temporal Ensembling
    1. Works under the assumption that **as the training progresses, averaging the predictions over time on unlabeled samples** may contribute to a better approximation of the true labels
    2. The network performs **the EMA (exponential moving average) to update the generated targets** @ every epoch (below figure is EMA)

$$f'(x)_t = \alpha f'(x)_{t-1} + (1-\alpha) f(x)_t$$



Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results

Antti Tarvainen
The Curious AI Company
and Aalto University
antti.tarvainen@aalto.fi

Harri Valpola
The Curious AI Company
harri@cai.fi

Related Work

2. Mean teacher
    1. Self-ensembling was extended to **directly combine model weights instead of predictions** => the Mean Teacher
    2. Updates the model weights at each step => generating a slightly improved model compared to the model w/o the EMA
    3. In this scenario, **EMA model was named teacher**, and **the standard model, student**

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$$

Related Work

$$\theta'_t = \alpha\theta'_{t-1} + (1 - \alpha)\theta_t$$

- theta : model parameters
- t : step
- alpha : **hyperparameter regulating the importance of the current model's weights w.r.t previous models (if alpha == 1 => same weight as previous step)**
- best result : alpha is increased later on during training

Related Work

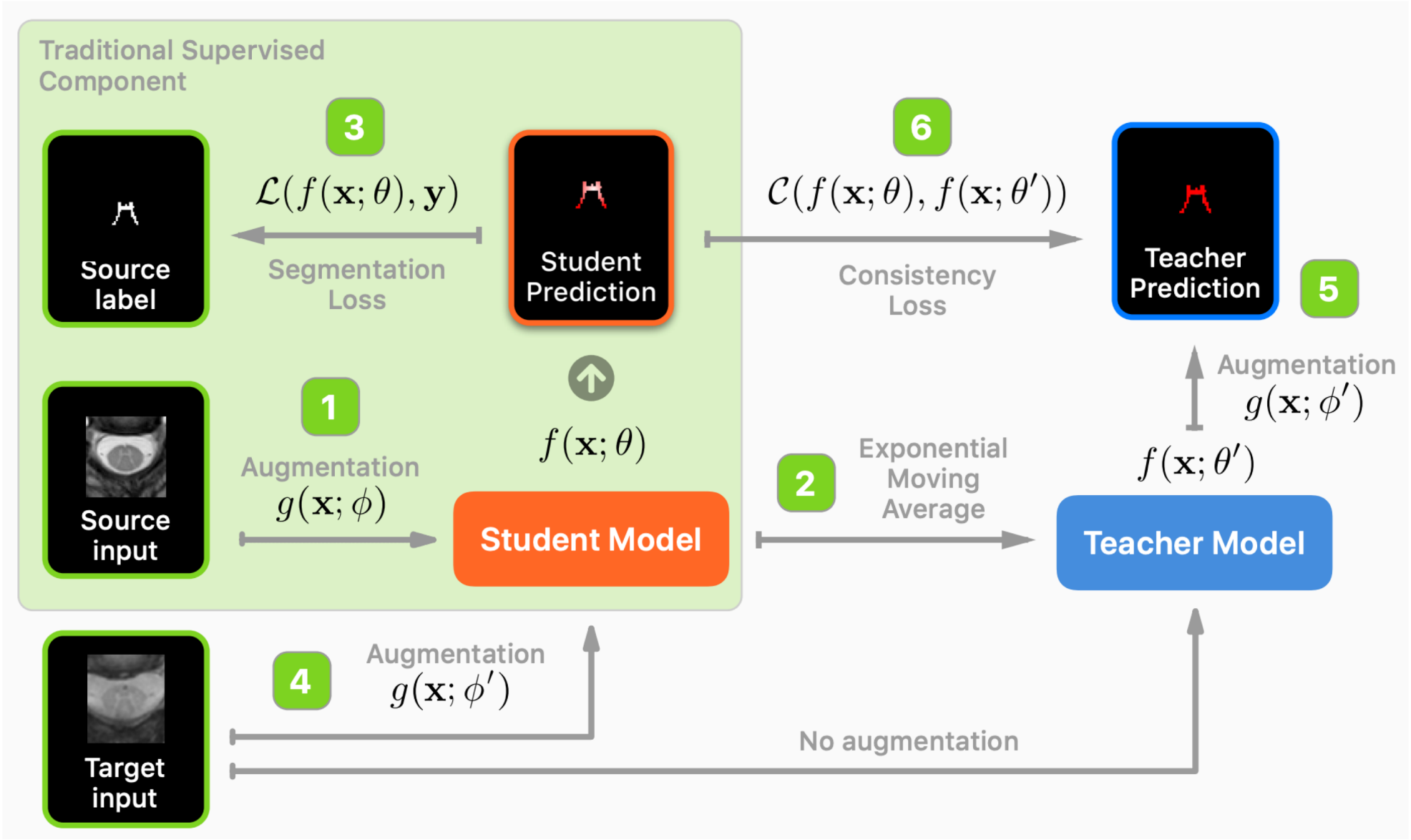
3. **Loss** for Mean teacher

$$J(\theta) = J_{task}(\theta) + \gamma J_{consistency}(\theta) + \lambda R(\theta)$$

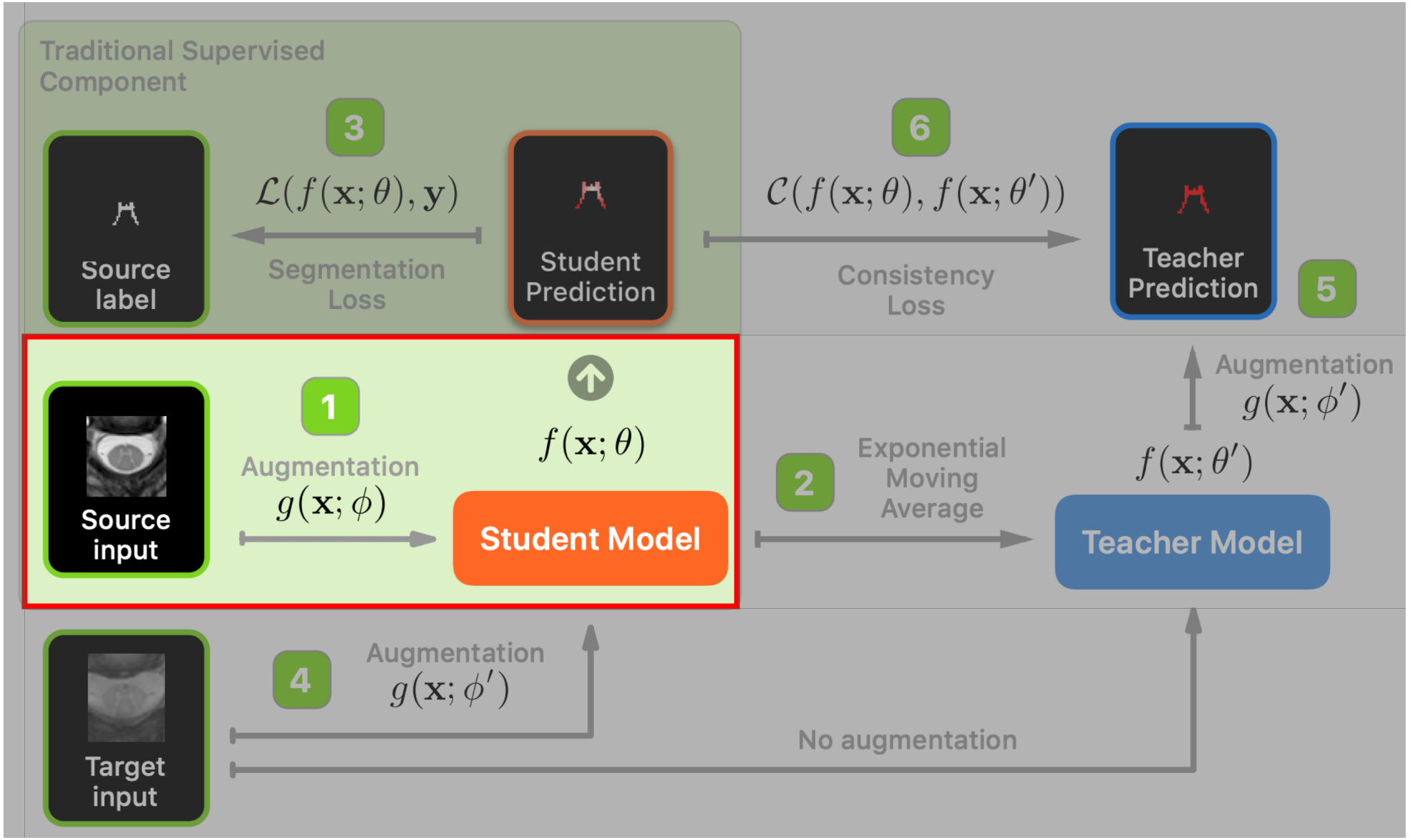
- **Gamma** : consistency weights (empirically found)
- **lambda** : regularization weights

1. Loss for both labeled and unlabeled data
2. All samples from a batch are **evaluated by both the student and teacher models**
    1. their **respective predictions compared via the consistency loss**
    2. labeled data is compared with ground truth (task loss)

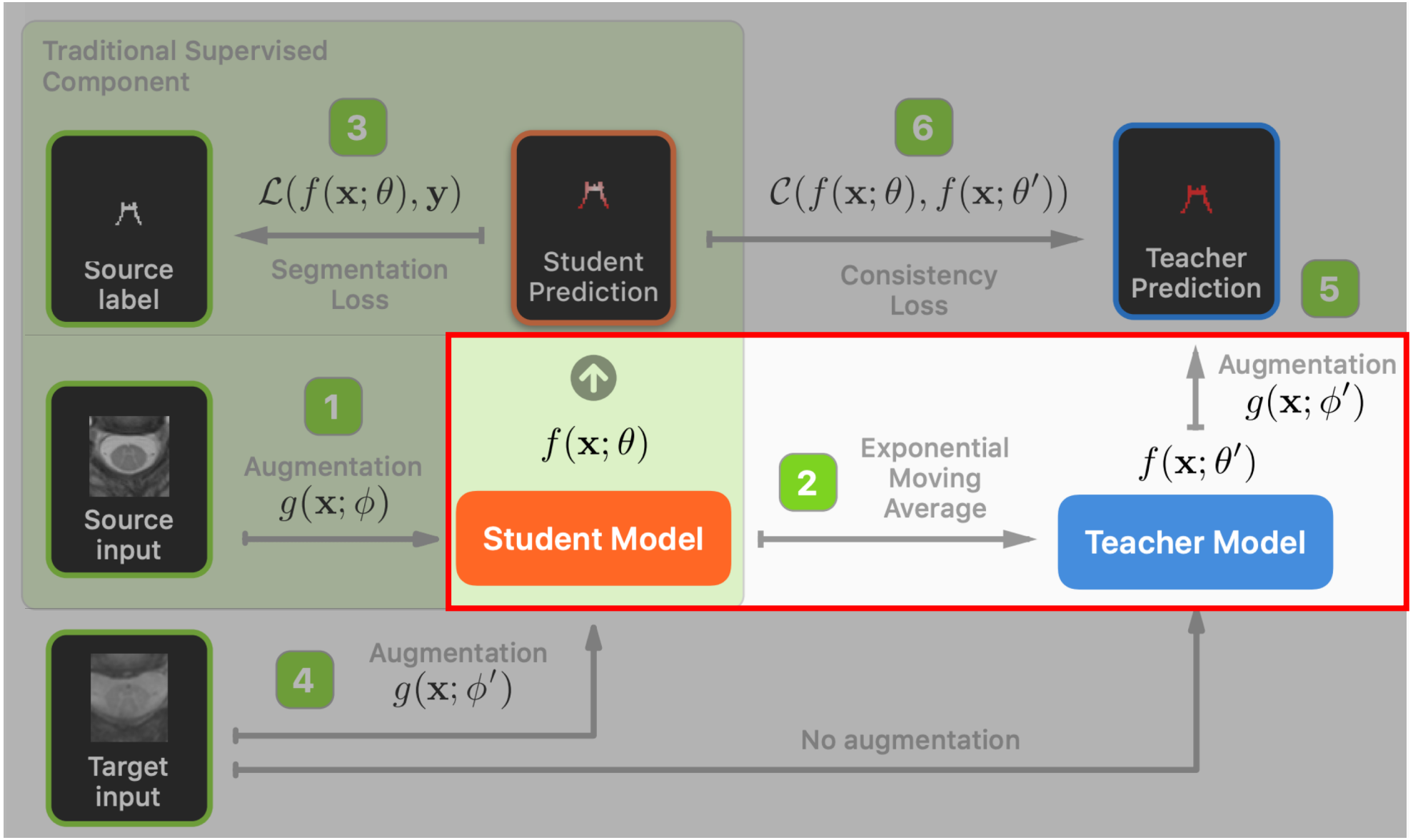
Materials & Methods



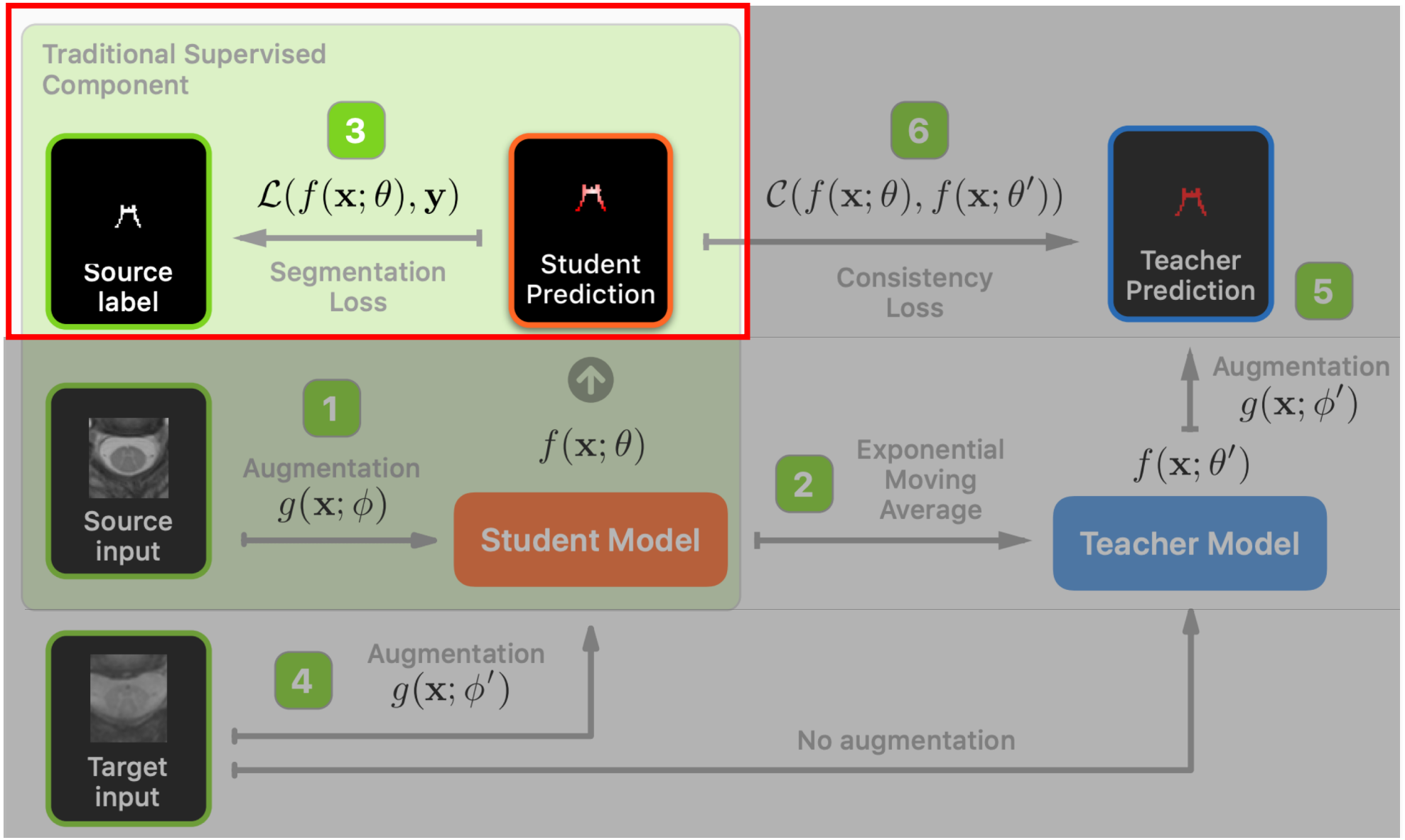
Materials & Methods



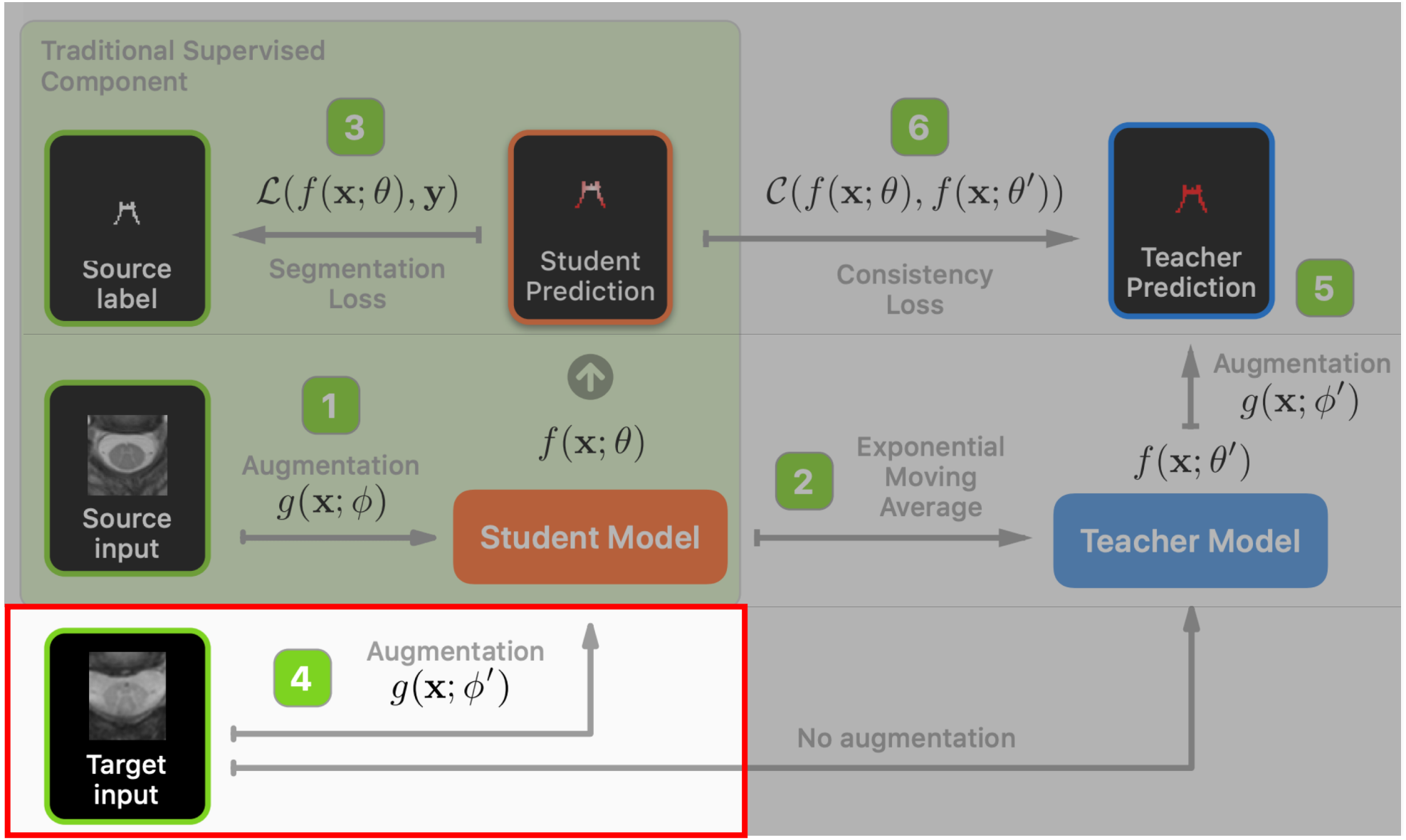
Materials & Methods



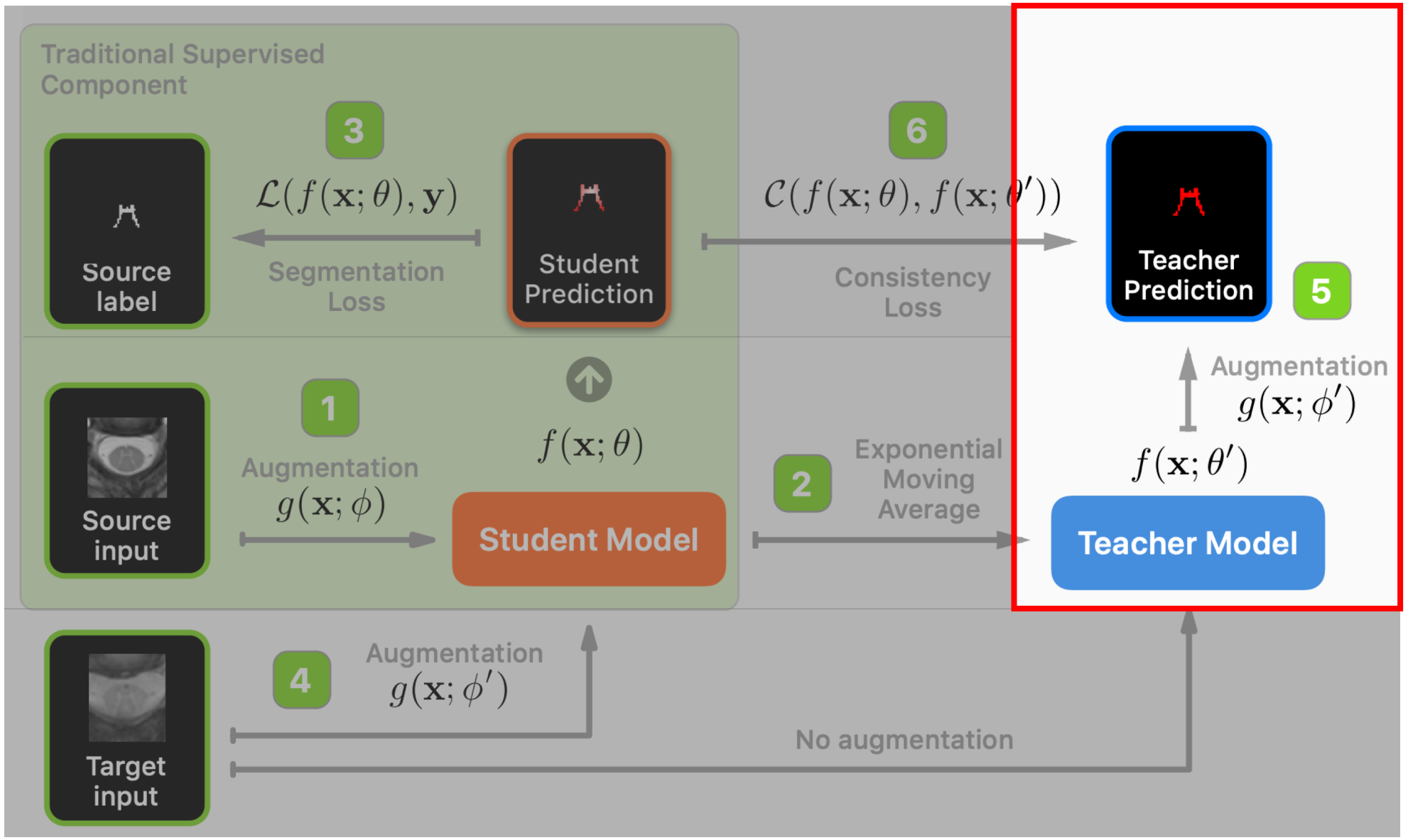
Materials & Methods



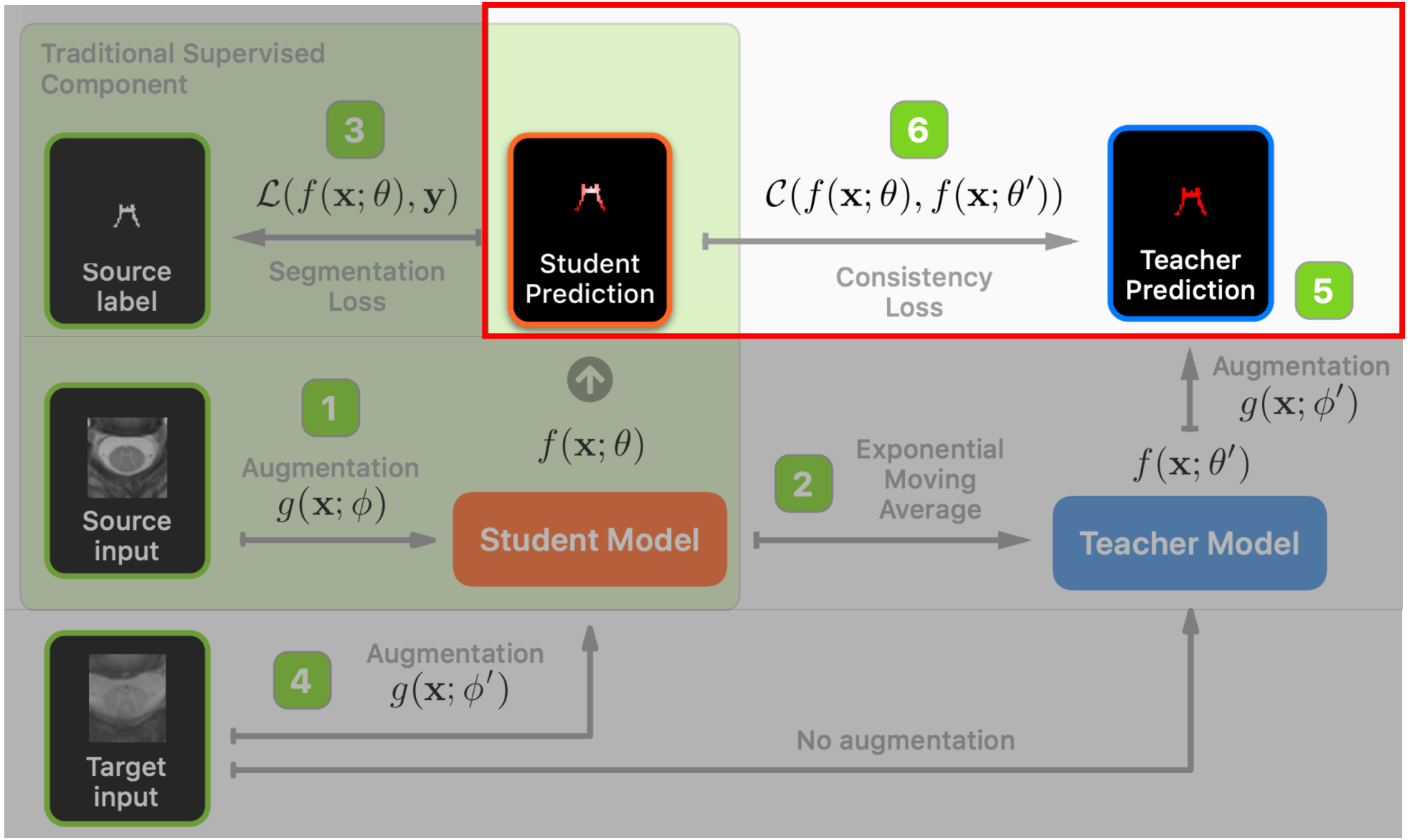
Materials & Methods



Materials & Methods



Materials & Methods



Adapting mean teacher for segmentation tasks

1. Task Loss : **Dice Loss** for segmentation task
    1. **Dice loss** was kept as the task loss for both baseline and adaptation experiments
    2. the dice loss is computed for the entire batch at once
2. Inconsistency btw training samples with spatial information
    1. Those augmentation parameters to both inputs are different
    2. To solve this, **spatial transformation** $g(x;\varphi)$ is applied to both student and teacher models
    3. There is no need for differentiation on the delayed augmentation on the teacher model (**Backprop only occurs for the student model**)

Model Architecture

1. U-Net
    1. 15 layers
    2. **group normalization**
        1. NOT batch normalization
        2. Discussed later
    3. Dropout
2. Mean Teacher model
    1. Also acts as a **regularizer** (kept the same regularization weights for all comparisons)

Baseline employed

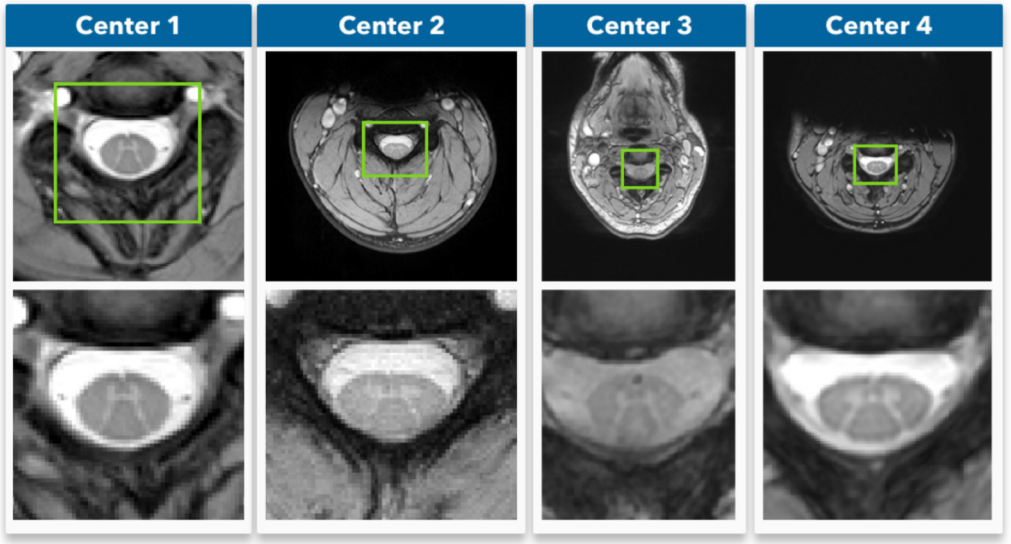
1. Hyperparameter Setting
    1. Minibatch size : 12
    2. dropout rate : 0.5
    3. Optimizer : Adam ([0.99, 0.999])
    4. learning rate : **Sigmoid learning rate ramp-up** strategy until epoch 50 + **cosine ramp-down** until epoch 350
2. No hyperparameter from the baseline model was changed in the adaptation scenario

Consistency loss - Important aspects of Mean Teacher

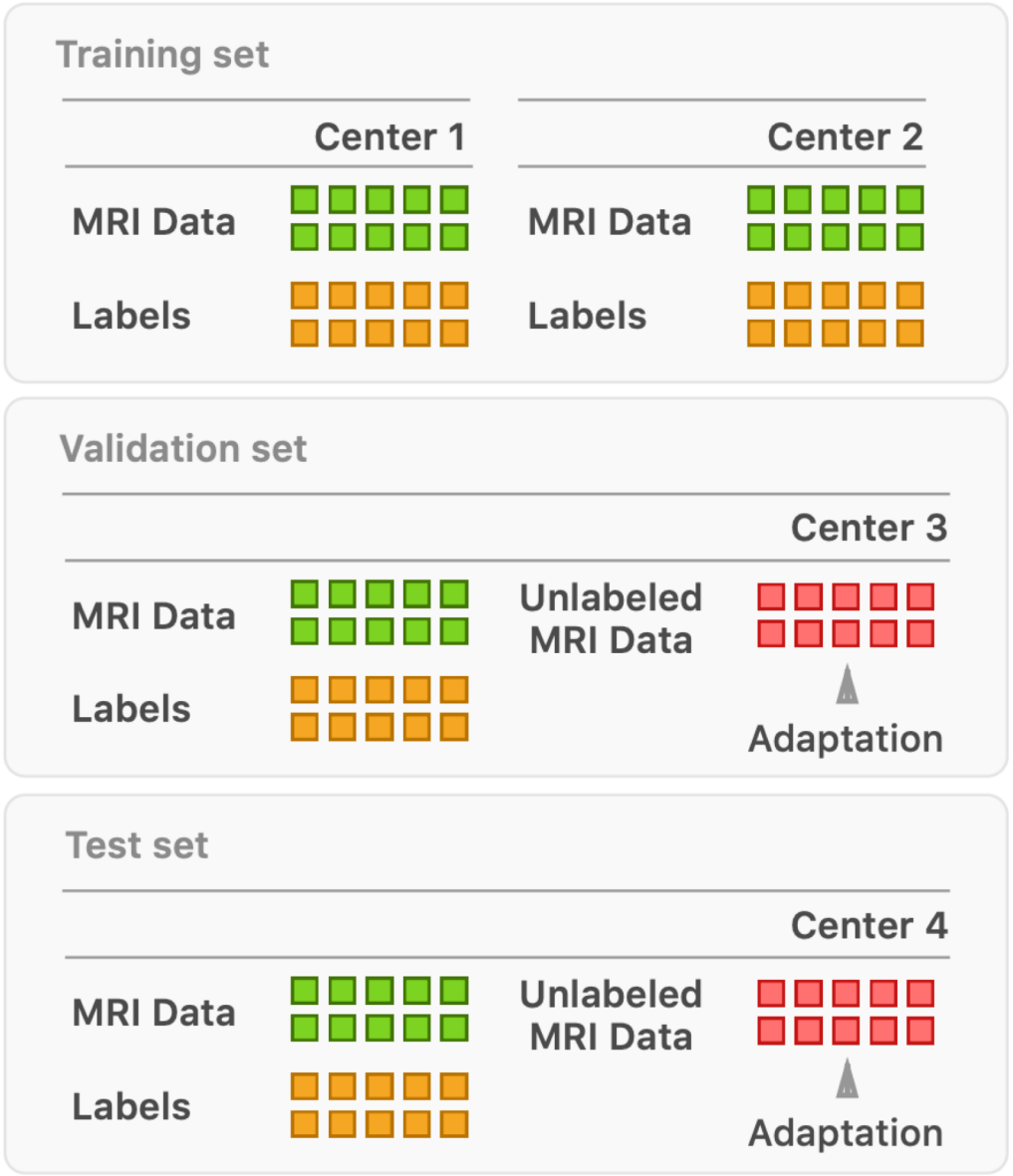
1. Student model - Teacher model loss
    1. Segmentation **relies on thresholding predictions from the teacher to define binary expected voxel values for the student**
    2. Focal loss, Dice loss, Tversky loss
        1. TV loss - Variation of Dice loss
        2. However **TV loss has many hyperparameters to determine**
    3. The output of teacher is soft (not binary)
        1. Dice, TV loss are not proper to be applied
2. Alternative thresholding method to **modify the formulations of the loss functions s.t. they can properly handle non-binary labels**

Materials & Methods

1. Spinal Cord Gray Matter Challenge dataset
    1. **multi-center**
    2. **multi-vendor**
    3. Three different MRI systems - **Philips, Siemens Trio and Skyra**
    4. 80 control / 20 subjects per each center
    5. Labels
        1. 4 gold standard segmentation masks / manually created by 4 independent experts
    6. Dataset Split
        1. Training - center 1 & 2
        2. Validation - center 3
        3. Test - center 4



Materials & Methods



Results

Table 1
Evaluation results in different centers. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. Results are averages and standard deviations over 10 runs (with independent initialization of random weights). Values highlighted represent the best results at each center. All experiments were trained in both centers 1 and 2 simultaneously. Dice represents the Sørensen–Dice coefficient and mIoU represents the mean Intersection over Union.

Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 1	Baseline	47.25 ± 0.10	31.46 ± 0.08	**94.90** ± 0.29	32.18 ± 0.09	99.66 ± 0.0	2.88 ± 0.01
	Center 3	47.71 ± 0.16	31.84 ± 0.14	94.18 ± 0.16	32.69 ± 0.15	99.67 ± 0.0	**2.85** ± 0.01
	Center 4	**48.42** ± 0.92	**32.47** ± 0.80	94.51 ± 0.57	**33.33** ± 0.93	**99.68** ± 0.02	2.86 ± 0.02
Center 2	Baseline	50.69 ± 0.09	34.44 ± 0.08	**94.79** ± 0.24	35.32 ± 0.10	99.61 ± 0.00	2.89 ± 0.01
	Center 3	51.05 ± 0.25	34.76 ± 0.23	93.78 ± 0.42	35.83 ± 0.31	99.62 ± 0.01	**2.87** ± 0.01
	Center 4	**51.29** ± 0.67	**34.98** ± 0.61	93.87 ± 0.91	**36.06** ± 0.82	**99.63** ± 0.02	2.87 ± 0.02
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61** ± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72** ± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17** ± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91** ± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22** ± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67** ± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

Results

Table 1

Evaluation results in different centers. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. Results are averages and standard deviations over 10 runs (with independent initialization of random weights). Values highlighted represent the best results at each center. All experiments were trained in both centers 1 and 2 simultaneously. Dice represents the Sørensen–Dice coefficient and mIoU represents the mean Intersection over Union.

Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 1	Baseline	47.25 ± 0.10	31.46 ± 0.08	**94.90** ± 0.29	32.18 ± 0.09	99.66 ± 0.0	2.88 ± 0.01
	Center 3	47.71 ± 0.16	31.84 ± 0.14	94.18 ± 0.16	32.69 ± 0.15	99.67 ± 0.0	**2.85** ± 0.01
	Center 4	**48.42** ± 0.92	**32.47** ± 0.80	94.51 ± 0.57	**33.33** ± 0.93	**99.68** ± 0.02	2.86 ± 0.02
Center 2	Baseline	50.69 ± 0.09	34.44 ± 0.08	**94.79** ± 0.24	35.32 ± 0.10	99.61 ± 0.00	2.89 ± 0.01
	Center 3	51.05 ± 0.25	34.76 ± 0.23	93.78 ± 0.42	35.83 ± 0.31	99.62 ± 0.01	**2.87** ± 0.01
	Center 4	**51.29** ± 0.67	**34.98** ± 0.61	93.87 ± 0.91	**36.06** ± 0.82	**99.63** ± 0.02	2.87 ± 0.02
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61** ± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72** ± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17** ± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91** ± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22** ± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67** ± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

Results

Table 1

Evaluation results in different centers. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. Results are averages and standard deviations over 10 runs (with independent initialization of random weights). Values highlighted represent the best results at each center. All experiments were trained in both centers 1 and 2 simultaneously. Dice represents the Sørensen–Dice coefficient and mIoU represents the mean Intersection over Union.

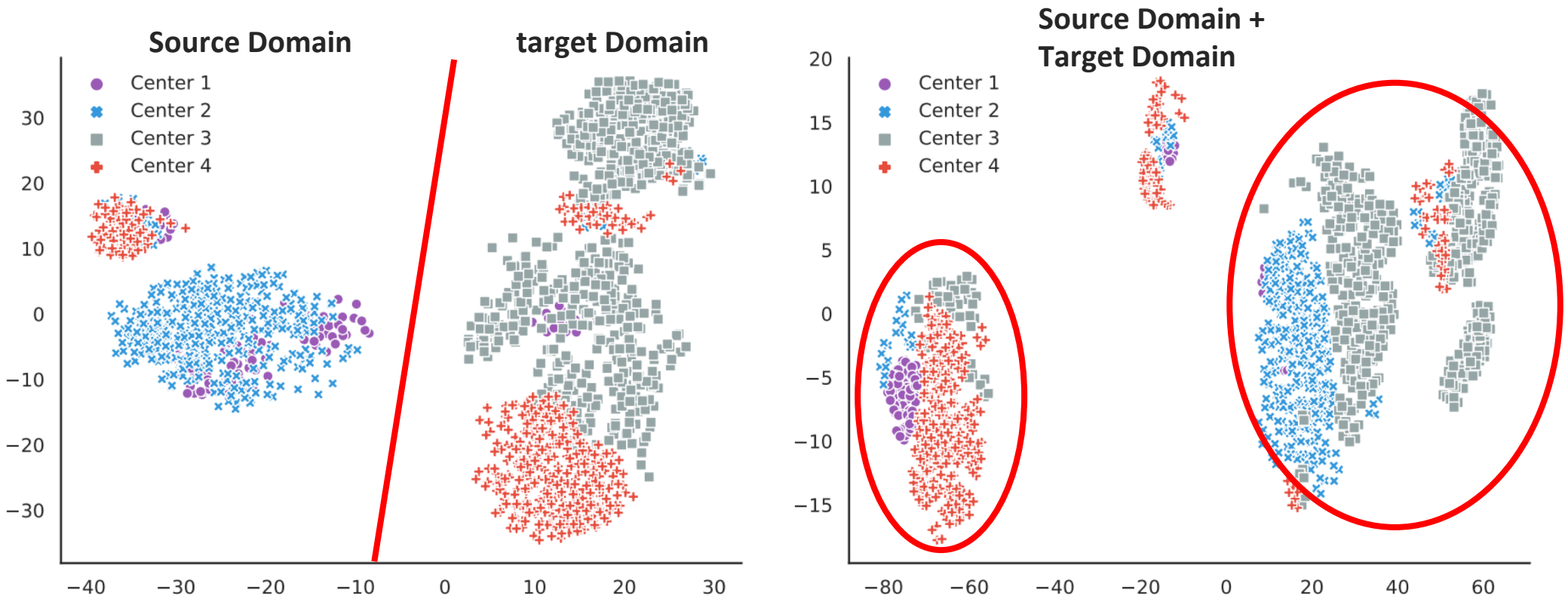
Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 1	Baseline	47.25 ± 0.10	31.46 ± 0.08	**94.90** ± 0.29	32.18 ± 0.09	99.66 ± 0.0	2.88 ± 0.01
	Center 3	47.71 ± 0.16	31.84 ± 0.14	94.18 ± 0.16	32.69 ± 0.15	99.67 ± 0.0	**2.85** ± 0.01
	Center 4	**48.42** ± 0.92	**32.47** ± 0.80	94.51 ± 0.57	**33.33** ± 0.93	**99.68** ± 0.02	2.86 ± 0.02
Center 2	Baseline	50.69 ± 0.09	34.44 ± 0.08	**94.79** ± 0.24	35.32 ± 0.10	99.61 ± 0.00	2.89 ± 0.01
	Center 3	51.05 ± 0.25	34.76 ± 0.23	93.78 ± 0.42	35.83 ± 0.31	99.62 ± 0.01	**2.87** ± 0.01
	Center 4	**51.29** ± 0.67	**34.98** ± 0.61	93.87 ± 0.91	**36.06** ± 0.82	**99.63** ± 0.02	2.87 ± 0.02
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61** ± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72** ± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17** ± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91** ± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22** ± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67** ± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

Results

Table 2
Results on evaluating on center 3. The training set includes centers 1 and 2 simultaneously, with unsupervised adaptation for center 3. Values within parentheses represent the best validation results for each metric. The remaining values represent the final result after 350 epochs.

Loss	Weight	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
CE	5	0.00 (85.50)	0.00 (74.91)	0.00 (95.01)	0.00 (98.90)	100.0 (100.00)	0.00 (0.00)
	10	0.00 (80.73)	0.00 (69.54)	0.00 (83.21)	0.00 (98.78)	100.0 (100.00)	0.00 (0.00)
	15	6.43 (37.03)	4.89 (26.06)	5.38 (77.05)	17.34 (65.85)	100.0 (100.00)	0.28 (0.00)
	20	2.30 (67.61)	1.86 (52.55)	2.09 (65.00)	7.94 (96.57)	100.0 (100.00)	0.12 (0.03)
Dice	5	76.76 (80.74)	62.76 (68.16)	97.88 (99.66)	63.72 (72.50)	99.71 (99.81)	2.36 (2.16)
	10	4.77 (10.55)	2.45 (5.64)	96.25 (99.99)	2.45 (5.85)	79.59 (99.75)	8.80 (2.57)
	15	2.30 (7.74)	1.16 (4.12)	99.95 (100.00)	1.16 (4.62)	55.07 (99.80)	11.75 (2.50)
	20	1.79 (4.43)	0.90 (2.27)	99.99 (100.00)	0.90 (2.30)	42.02 (99.84)	12.68 (2.43)
MSE	5	83.7 (83.88)	72.2 (72.46)	91.24 (98.19)	78.1 (78.57)	99.87 (99.93)	2.1 (2.00)
	10	84.38 (84.38)	73.19 (73.19)	90.15 (99.07)	80.12 (80.12)	99.88 (99.94)	2.05 (1.89)
	15	84.59 (84.59)	73.49 (73.50)	89.19 (98.52)	81.28 (81.28)	99.89 (99.89)	2.03 (2.03)
	20	84.5 (84.50)	73.36 (73.37)	90.36 (94.63)	80.16 (80.16)	99.88 (99.98)	2.05 (1.46)

Results - Visualization



(a) A visualization of the t-SNE 2D non-linear embedding projection for the supervised learning scenario. The colors represent data from different centers.

(b) A visualization of the t-SNE 2D non-linear embedding projection for the domain adaptation scenario. The colors represent data from different centers.

Figure 8. Execution of t-SNE algorithm for two different scenarios. Best viewed in color.

Conclusion & Limitation

1. Conclusion
    1. Showed that unsupervised DA, without depending on annotations, is an effective way to increase the performance of machine learning models
    2. Showed **how self-ensembling methods can improve generalization on unseen domains** through **the leverage of unlabeled data from multiple domains**
2. Limitation
    1. Didn't evaluate **adversarial training methods** for Domain adaptation
    2. Single-tast evaluation of the gray matter segmentation could be extended to other tasks

Thank you