MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data

Quande Liu, Qi Dou, Lequan Yu and Pheng Ann Heng, 2020

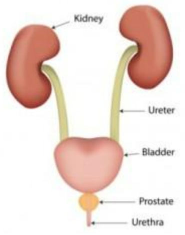
Kobi Rahimi

Bar Ilan University

Outline

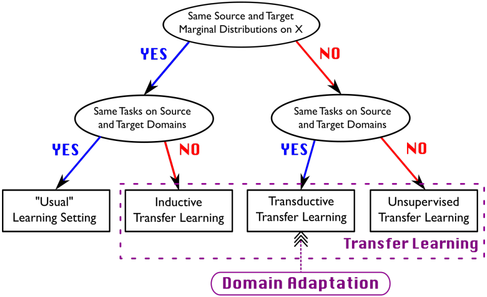
1. **Introduction**
   - Prostate Segmentation
   - Domain Shift
2. **Multi-Site Learning**
   - Motivation
   - Challenges
   - Approach
   - Methodology
3. **Learning Procedure**
   - Loss functions
   - Training procedure
4. **Experiments & Results**
   - Dataset
   - Experiments & Results

Prostate Segmentation



- The prostate is a gland and reproductive organ found in males.
- Prostate diseases are common afflictions in men.
- Accurate segmentation of the prostate from MRI images is crucial for diagnosis and treatment planning of these diseases.

Domain Shift



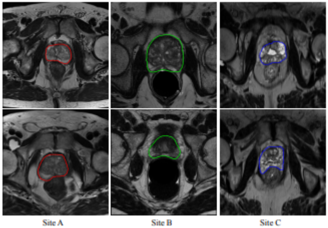
A domain shift, is a change in the data distribution between an algorithm's training dataset (source), and a dataset it encounters when deployed (target).

Motivation

- Due to the nature of scarcity for medical images, it is important to effectively aggregate data from multiple sites for robust model training.
- Jointly learn a single network from S different sites.
- Goal: boost the segmentation performance on all the S sites consistently.

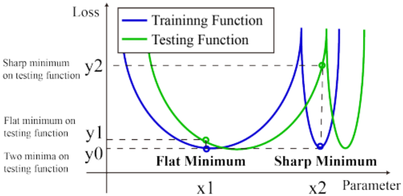
First challenge - heterogeneous datasets

- Prostate MRIs from different sites present heterogeneity.
  - e.g, differences in scanners, imaging protocols, etc.
- Causes difficulty for learning generic representations.
- The BN layers may result in inaccurate estimation of global mean and variance leading to performance degradation.



Second challenge - robust representation

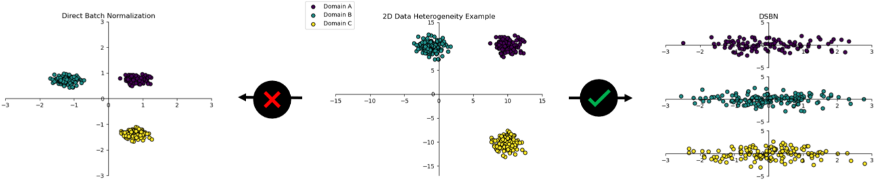
- Flat vs. sharp minima:



- Under multi-site learning, the model converges to sharp minima of each dataset instead of exploring a robust global minima among multiple datasets.
- Shifting in test (target) loss causes bad generalization.

Domain-Specific Batch Normalization (DSBN)

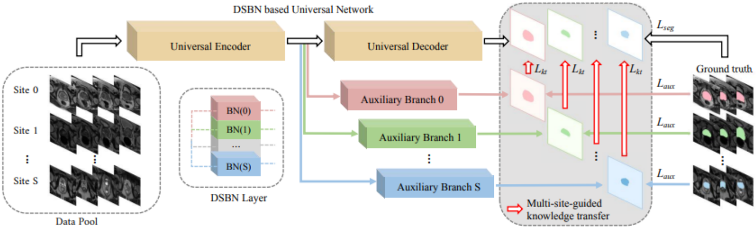
- DSBN layers are implemented to compensate for the inter-site heterogeneity of different MRI datasets.
- Enables the network to estimate statistics and perform feature normalization for each site separately.



Multi-site-guided Knowledge Transfer (MSKT)

- MSKT is proposed in order to capture the shared knowledge from multiple datasets.
- Enhances the CNN kernels to extract more generic representations from multi-site data by digging the flat minima among different datasets.

Methodology



- Universal decoder and auxiliary branches have the same architecture.
- Universal decoder substitutes the BN layers with DSBN layers.
- Training is done with dice loss.

DSBN Layer

- Assign an individual BN layer for each site (s).
- Normalize the internal representations along the **channel** dimension.

Mathematical Expression:

- Let $x_k^s \in [x_1^s, ..., x_K^s]$ be a certain channel of the K-channel feature maps in a certain layer, the corresponding normalized representation $y_k^s \in [y_1^s, ..., y_K^s]$ are computed as:

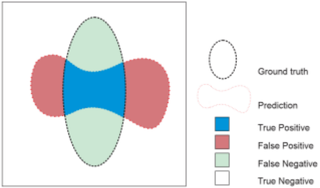
$y_k^s = \gamma^s \cdot \hat{x}_k^s + \beta^s$, where $\hat{x}_k^s = \dfrac{x_k^s - E[x_k^s]}{\sqrt{Var[x_k^s] + \epsilon}}$

MSKT via auxiliary branches

- Each auxiliary branch learns specific knowledge from a specific site.
- Train the universal network with dice loss w.r.t auxiliary branch's output (L_{kt} = 1 - dice score).

Suggested dice score:

$$\frac{2 \cdot Intersection}{Groundtruth^2 + Prediction^2} = \frac{2 \cdot TP^2}{(TP + FN)^2 + (TP + FP)^2}$$



MSKT via auxiliary branches (cont.)

- The universal network outputs a probability map of the predicted segmentation (M^s_{uni}).
- The auxiliary branches outputs a segmentation mask (P^s_{aux}).
- Knowledge transfer uses dice loss to align results from universal network and auxiliary branches.

Knowledge transfer dice loss:

$$L^s_{kt}(M^s_{uni}, P^s_{uni}) = 1 - \frac{2\sum_i^{\Omega} m^s_i \cdot p^s_i}{\sum_i^{\Omega}(m^s_i)^2 + \sum_i^{\Omega}(p^s_i)^2}$$

- Ω - total number of pixels in one batch.
- $m^s_i \in M^s_{uni}, p^s_i \in P^s_{aux}$.

Loss functions

Auxiliary branches:

$$L_{aux} = \sum_{s=1}^{S} L_{aux}^{s} + \eta(||\theta_e||_2^2 + \sum_{s=1}^{S} ||\theta_{aux}^{s}||_2^2)$$

Universal network:

$$L_{uni} = \sum_{s=1}^{S} \alpha L_{kt}^{s} + (1-\alpha) L_{uni}^{s} + \eta(||\theta_e||_2^2 + ||\theta_d||_2^2)$$

- L_{aux}^{s}, L_{uni}^{s} - Segmentation dice loss for auxiliary branches and universal network w.r.t ground truth labels.
- η - L_2 regularization factor ($1e^{-4}$).
- α - Balances between segmentation and knowledge transfer losses (0.5).
- $\theta_e, \theta_d, \theta_{aux}^{s}$ - Universal encoder, decoder and auxiliary parameters.

Training procedure

Algorithm 1: Training procedure of the proposed MS-Net

Data: Datasets $D_1, \ldots, D_S$ from from S different sites,
Training iteration τ

Result: Universal encoder $\boldsymbol{\theta}_e$, Universal decoder $\boldsymbol{\theta}_d$,
Auxiliary branch $\{\boldsymbol{\theta}_i\}_{i=1}^S$;

—**Training**—

Initialization: t=1; Randomly initialize $\boldsymbol{\theta}_e$,$\boldsymbol{\theta}_d$,$\{\boldsymbol{\theta}_i\}_{i=1}^S$;

while $t \leq \tau$ **do**

Given S mini-batches from $\{D_i\}_{i=1}^S$;
Compute the loss function of all auxiliary branches L_{aux};
Update parameters $\boldsymbol{\theta}_e$ and $\{\boldsymbol{\theta}_i\}_{i=1}^S$;
Compute the loss function of universal network L_{uni};
Update parameters $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_d$;

end

—**Testing**—

Leave $\{\boldsymbol{\theta}_i\}_{i=1}^S$ and only keep $\boldsymbol{\theta}_e$, $\boldsymbol{\theta}_d$ for deployment.

Dataset

Dataset	Case num	Field strength (T)	Resolution(in-plane/through-plane)(mm)	Coil	Manufactor
Site A	30	3	0.6-0.625/3.6-4	Surface	Siemens
Site B	30	1.5	0.4/3	Endorectal	Philips
Site C	19	3	0.67-0.79/1.25	No	Siemens

- **Site A** - Radboud University Nijmegen Medical Center, 30 samples.
- **Site B** - Boston Medical Center, 30 samples.
- **Site C** - Hospital Center Regional University of Dijon-Bourgogne, 19 samples.

Separate vs. Joint experiment

Separate approach:

- Train a separate model for each dataset.
- Test within and across datasets.

Joint approach:

- Normalize each dataset separately before entering the model.
- Train one model with all the normalized datasets.
- Test with all the datasets.

Results

Methods	BFC	NF	Intensities	Site A	Site B	Site C	Overall
Separate (A)	✗	✗	whitening	90.47	76.44	56.81	
Separate (B)	✗	✗	whitening	70.11	90.52	50.25	90.56
Separate (C)	✗	✗	whitening	57.93	55.25	90.70	
Joint	✗	✗	✗	86.51	88.00	86.78	87.10
Joint	✗	✗	histogram	87.68	88.02	89.46	88.39
Joint	✗	✗	scaled	90.43	88.06	88.26	88.92
Joint	✗	✗	whitening	90.69	89.53	90.55	90.25
Joint	✗	✓	whitening	90.76	89.46	90.91	90.37
Joint	✓	✗	whitening	90.84	89.81	90.81	90.49
Joint	✓	✓	whitening	91.14	89.75	90.83	90.58

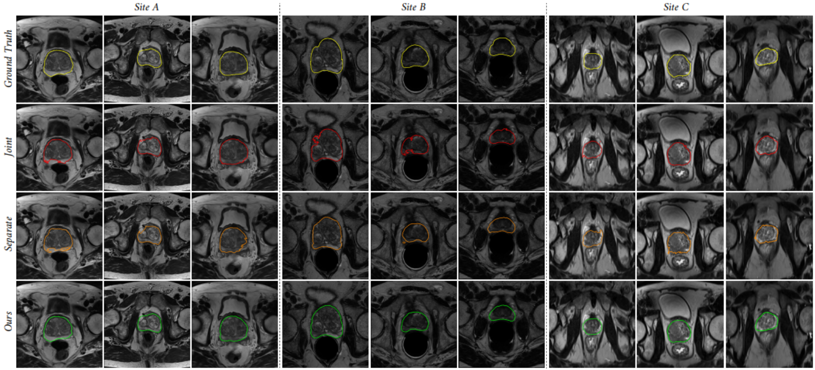
- Joint approach doesn't show explicit advantage over the separate approach in terms of accuracy.

MS-Net vs. other state of the art experiment

	Dice Coefficient (mean±std, %)				Average Symmetric Distance (mean±std, mm)			
Methods	Site A	Site B	Site C	Overall	Site A	Site B	Site C	Overall
Tian *et al.* [26]	88.23	88.23	—					
Rundo *et al.* [9]	—	—	88.66					
Separate	90.47±3.00	90.52±2.45	90.70±3.34	90.56±2.88	1.02±0.42	0.84±0.32	0.75±0.33	0.87±0.38
Joint	90.69±3.05	89.53±2.97	90.55±3.18	90.25±3.08	0.96±0.37	0.90±0.34	0.75±0.32	0.87±0.36
USE-Net [19]	90.90±2.41	90.17±2.61	90.73±2.36	90.60±2.50	0.90±0.32	0.85±0.31	0.70±0.30	0.82±0.32
Dual-Stream [47]	90.87±2.85	90.57±2.12	90.10±3.28	90.51±2.72	0.92±0.38	0.84±0.27	0.75±0.32	0.83±0.33
Series-Adapter [48]	90.80±2.72	89.92±2.80	91.24±2.21	90.65±2.71	0.95±0.42	0.92±0.38	0.71±0.28	0.86±0.39
Parallel-Adapter [23]	90.61±3.54	90.71±2.17	91.30±2.06	90.88±2.79	0.96±0.25	0.83±0.29	0.74±0.28	0.84±0.28
DSBN (ours)	90.98±2.69	90.67±2.22	91.07±1.86	90.91±2.36	0.95±0.48	0.83±0.30	0.74±0.24	0.84±0.38
MS-Net (ours)	**91.54±2.01**	**91.24±1.97**	**92.18±1.62**	**91.66±1.95**	**0.89±0.33**	**0.76±0.25**	**0.67±0.23**	**0.77±0.29**

- MS-Net outperforms each and every state of the art approach.

Segmentation Results



Questions?

Bibliography

- MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data. Quande Liu, Qi Dou, Lequan Yu and Pheng Ann Heng.
  https://arxiv.org/pdf/2002.03366.pdf
- Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, Kaisheng Ma.
  https://arxiv.org/pdf/1905.08094.pdf
- Alejandra Marquez Herrera, Alex Cuadros Vargas, Helio Pedrini. Improving Semantic Segmentation of 3D Medical Images on CNNs. LatinX in AI Research at ICML 2019, Jun 2019, California, United States. ffhal-02265952f
  https://hal.archives-ouvertes.fr/hal-02265952/document