国防科技大学
National University of Defense Technology

icassp 2022
Singapore

Adaptive Pseudo Labeling For Source-Free Domain Adaptation In Medical Image Segmentation

Chen Li†, Wei Chen†✉, Xin Luo , Yulin He, Yusong Tan

College of Computer, National University of Defense Technology, Changsha, China

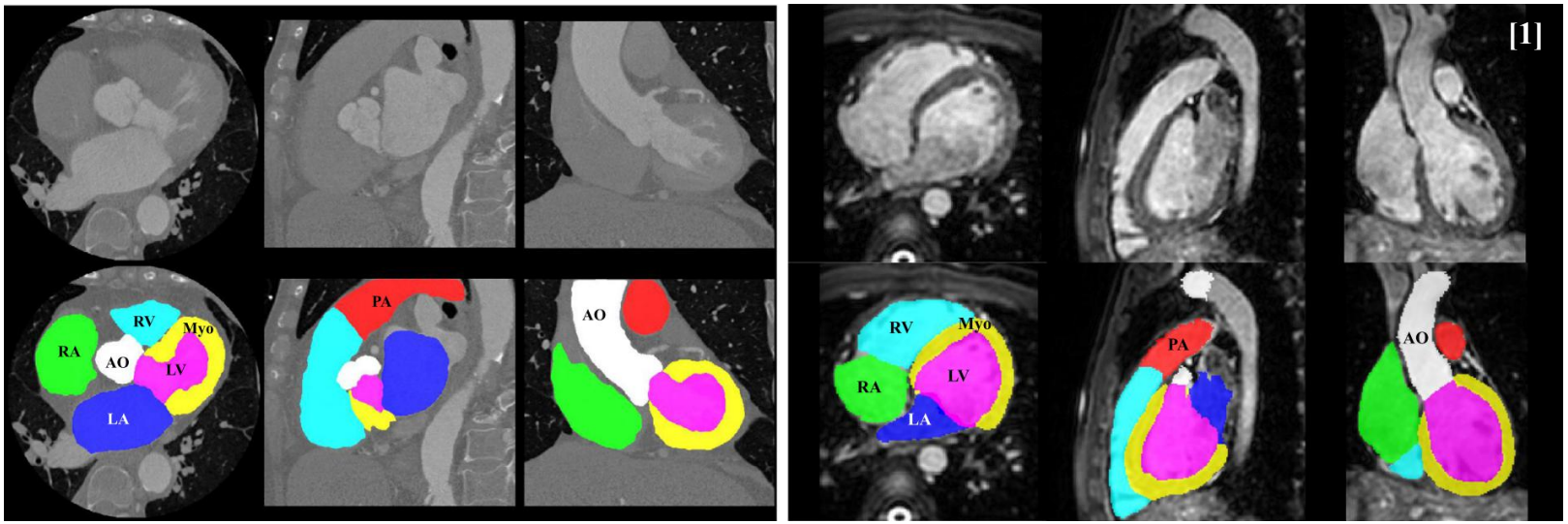
† These authors contributed equally to this work

✉ *chenwei@nudt.edu.edu*

ICASSP 2022, 7-13 May, Singapore(Virtual)

Medical Images Segmentation (MIS)

- Medical image segmentation means classifying pixel-wise segments into different components from biomedical data (CT, MRI, Ultrasound, cells scan ……)



- Medical image segmentation is an essential step and plays a crucial role in many clinical applications, such as disease diagnosis and treatment planning.
- Segmentation from medical images is more challenging than natural image.

[1] Zhuang, Xiahai, et al. "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge." Medical image analysis 58 (2019).

Limitations in supervised MIS methods

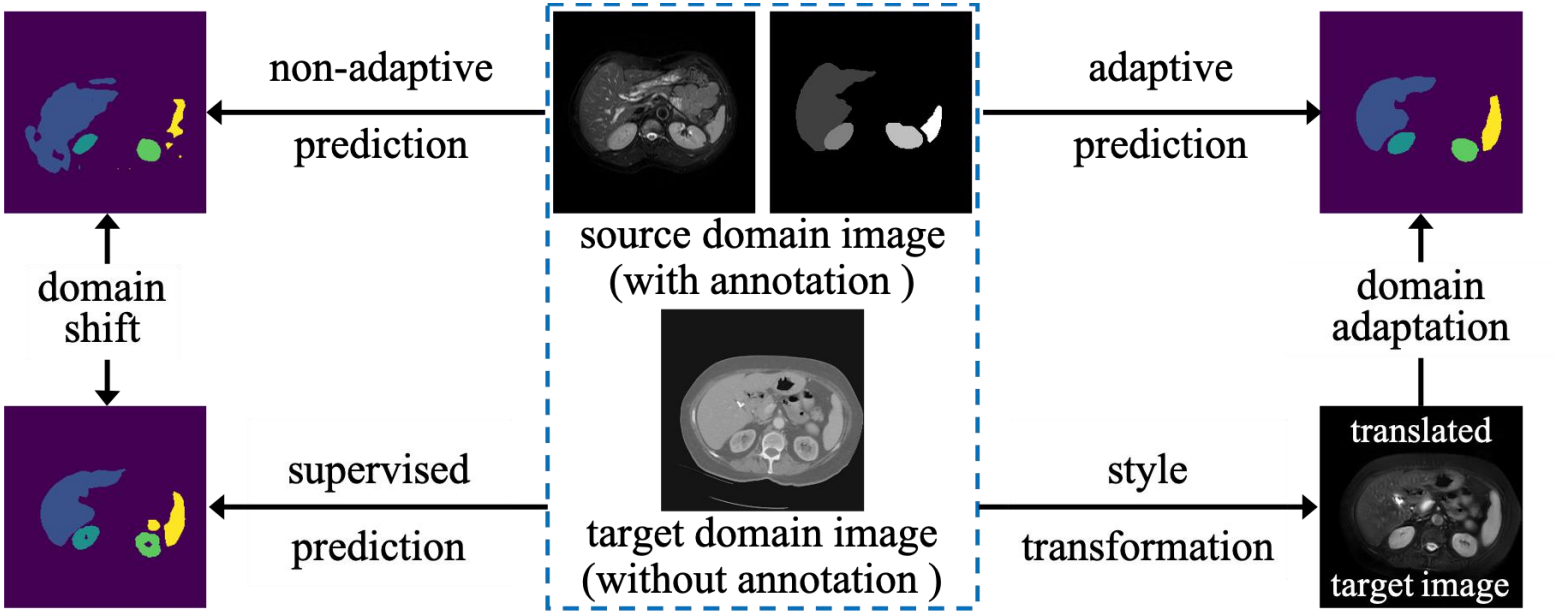
- Supervised methods have shown promising performances in various medical image segmentation tasks.

- Well-trained models often fail when deployed to real-world clinical scenarios, as medical images acquired with different acquisition parameters or modalities have very different characteristics.

- Such cross-modality domain shift would lead to severe performance degradation of deep networks.

Unsupervised Domain Adaptation (UDA)

- The main idea of UDA is to extract domain-invariable representations and transfer them from source domain to target domain, where source samples are annotated and the the labels of target samples are absent.



Motivation

- Existing works mostly still require manually design thresholds to separate high confident objects, which were not an automatic learning process and neglected confident pixels.
- The uncertainty regularization-based method ignored the inconsistency of different noisy predictions, and failed to take advantage of the complementary relationship between consistency and confidence.

Motivation

- We investigate two critical properties for the generated pseudo labels
  - Consistency is measured by calculating the variance of predictions from different branches and evaluated as the internal reliability of the pseudo label.
  - Confidence is measured by calculating the discrepancy of pseudo label and evaluated by the external reliability of the pseudo label.
  - Higher consistency and confidence denote better prediction.

A=[0.34,0.33,0.33]
B=[0.35,0.32,0.33]

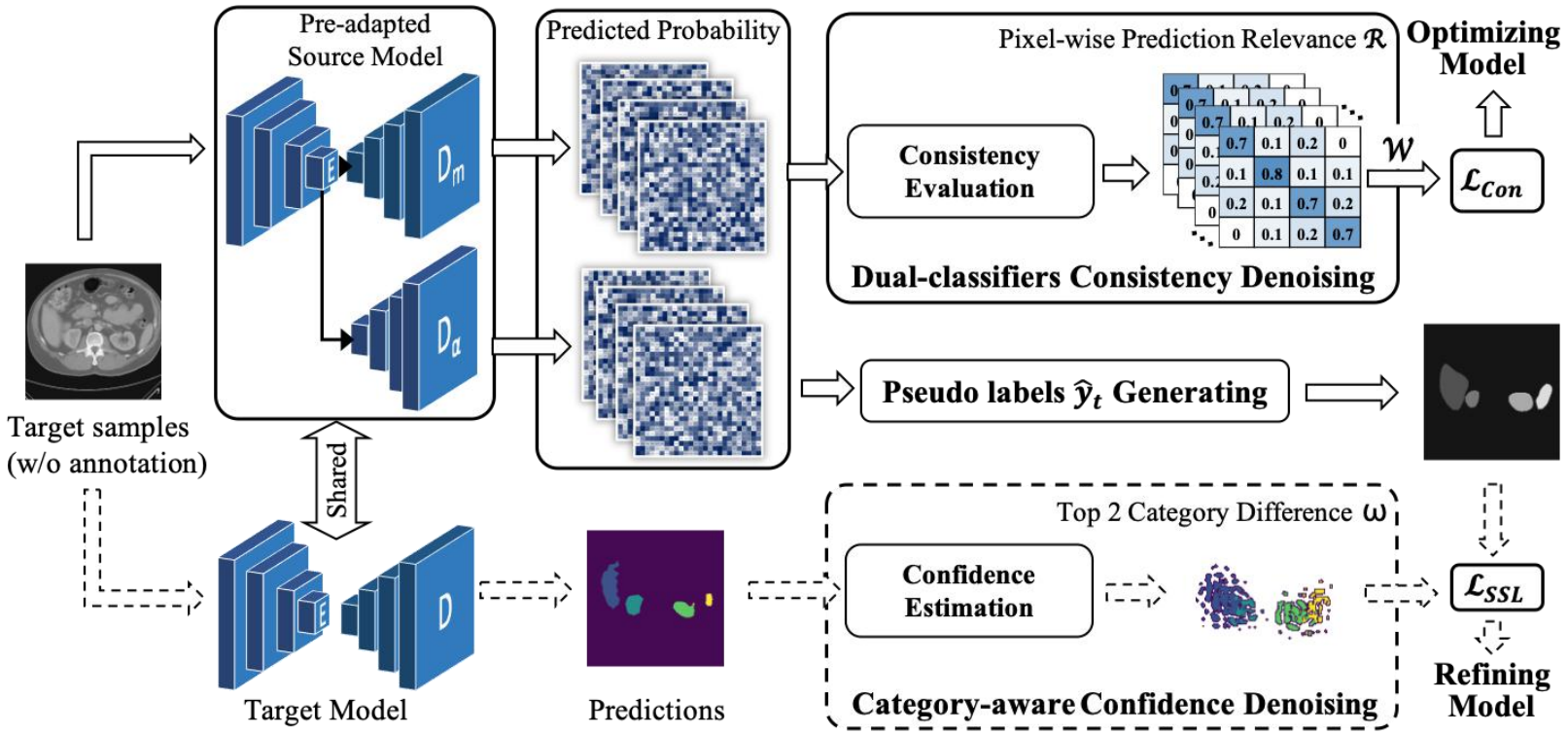
➔Predictions are consistent in distribution but not confident.

C= [0.80,0.15,0.05]
D=[0.70,0.15,0.15]

➔Predictions are confident but inconsistent.

Adaptive Pseudo Labeling framework

- We propose a full automatic **A**daptive **P**seudo-**L**abeling framework (APL).
- Dynamic pixel-wise weights replace manually designed thresholds.
- DeepLabv2 + dual-branch classifiers are selected as the baseline model.



Dual-Classifiers Consistency Denoising

- Referring to the BCDM[1], we redesign the loss function to measure the pixel-wise consistency between the prediction results dual-classifiers.
- We define the weighted Dual-Classifiers Prediction Relevance R to evaluate the relevance across different classes between the dual classifiers predictions.:

$$\mathcal{R}_{x_t} = \mathcal{W}_{x_t} \times D_m(E(x_t)) \times D_a(E(x_t))^T$$

$\mathcal{W}_{x_t}$ is a weight matrix with size C × C, where the value in the i-th row and j-th column is the number of pixels.

[1] *Shuang Li, et al. "Bi-classifier determinacy maxi mization for unsupervised domain adaptation," in AAAI Conference on Artificial Intelligence (AAAI), 2021.*

Dual-Classifiers Consistency Denoising

- Referring to the BCDM[1], we redesign the loss function to measure the pixel-wise consistency between the prediction results dual-classifiers.
- We define the weighted Dual-Classifiers Prediction Relevance R to evaluate the relevance across different classes between the dual classifiers predictions.:

$$\mathcal{R}_{x_t} = \mathcal{W}_{x_t} \times D_m(E(x_t)) \times D_a(E(x_t))^T$$

- The values of the diagonal element R determine the consistency of the pseudo labels.
- The larger the sum, the more concentrated the class distribution in the predictions and the higher the consistency.

[1] *Shuang Li, et al. "Bi-classifier determinacy maxi_x0002_mization for unsupervised domain adaptation," in AAAI Conference on Artificial Intelligence (AAAI), 2021.*

Dual-Classifiers Consistency Denoising

- With the help of Relevance R, the pseudo labels are denoised and consistent by maximizing the sum of the diagonal elements. Finally, we define the Dual classifiers Consistency loss on as the follows:

$$\min_{\theta_{D_m, D_a}} \mathcal{L}_{Con} = \sum_{x_t}^{X_T} [\sum_{i,j=1}^{C} \mathcal{R}_{x_t}^{i,j} - \sum_{i=1}^{C} \mathcal{R}_{x_t}^{i,i}]$$

Category-aware Confidence Denoising

- The confidence of each pseudo label is estimated by the difference between the probability values of the top two categories.
- Firstly, we merge the dual-classifier predictions and obtain the pseudo label as follows.

$$\hat{y}_t = arg\ max[D_m(E(x_t)) + D_a(E(x_t))]$$

- Category-aware Confifidence is calculated and formulated as follows.

$$\omega = \left|\delta_1(\hat{y}_t) - \delta_2(\hat{y}_t)\right|$$

where ω is the pixel-wise measured confidence, δ1 denotes the largest category probability among the pseudo label yˆt. In the same way, δ2 is the second largest category probability.

Category-aware Confidence Denoising

- Based on the denoised labels, we directly utilize them as the original target domain dataset's annotation and conduct Semi-Supervised Learning to refine the adapted model.

$$\min_{\theta_E} \mathcal{L}_{SSL} = \sum_{x_t \in \mathbb{X}_T, \hat{y}_t \in \hat{\mathbb{Y}_T}} \omega \cdot [-\hat{y}_t \log D(E(x_t))]$$

Dataset & Metric

- Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS)
  - 120 DICOM data sets from two different MRI sequences
  - supported by ISBI 2019
  - https://chaos.grand-challenge.org/Data/

- Multi-Atlas Labeling Beyond the Cranial Vault(MALBCV)
  - 30 labeled volumes from the CT training dataset.
  - supported by MICCAI 2015
  - https://www.synapse.org/Synapse:syn3193805/wiki/217789

- Evaluation Metrics:
  - Dice coefficient

Quantitatively measure the domain shift

- Firstly, we quantitatively measured the domain shift by calculating performance dependency in the abdominal organs segmentation.

Dice (%)	W/o Adapt	Supervised Train
liver	73.1	92.8
right kidney	47.3	86.4
left kidney	57.3	87.4
spleen	55.1	88.2
Avg	58.2	88.7

Comparison with the State-of-the-art UDA Methods

- Second, in order to evaluate the promotion of adaptive pseudo labeling method based on the adapted models, we applied the proposed APL method to other seven UDA methods to segment multi-abdominal organs from cross-modality medical images.

	UDA Methods															
Dice (%)	W/o Adapt	Supervised Train	SynSeg Net[1]	SynSeg Net[1]+APL	AdaOutput[2]	AdaOutput[2]+APL	CycleGAN[3]	CycleGAN[3]+APL	CyCADA[4]	CyCADA[4]+APL	ADVENT[5]	ADVENT[5]+APL	SIFA-v1[6]	SIFA-v1[6]+APL	SIFA-v2[7]	SIFA-v2[7]+APL
liver	73.1	92.8	85	85.19	85.4	87.65	83.4	86.79	84.5	89.06	89.28	90.37	87.9	89.56	88	89.56
right kidney	47.3	86.4	82.1	81.26	79.7	83.85	79.3	78.54	78.6	81.43	77.05	85.1	83.7	89.74	83.3	87.88
left kidney	57.3	87.4	72.7	79.24	79.7	83.88	79.4	79.24	80.3	83.3	81.37	81.48	80.1	83.66	80.9	89.19
spleen	55.1	88.2	81	84.41	81.7	80.77	77.3	84.62	76.9	81.33	83.45	84.82	80.5	82.31	82.6	82.8
Avg	58.2	88.7	80.2	82.53	81.6	84.04	79.9	82.3	80.1	83.78	82.79	85.44	83.1	86.32	83.7	87.11

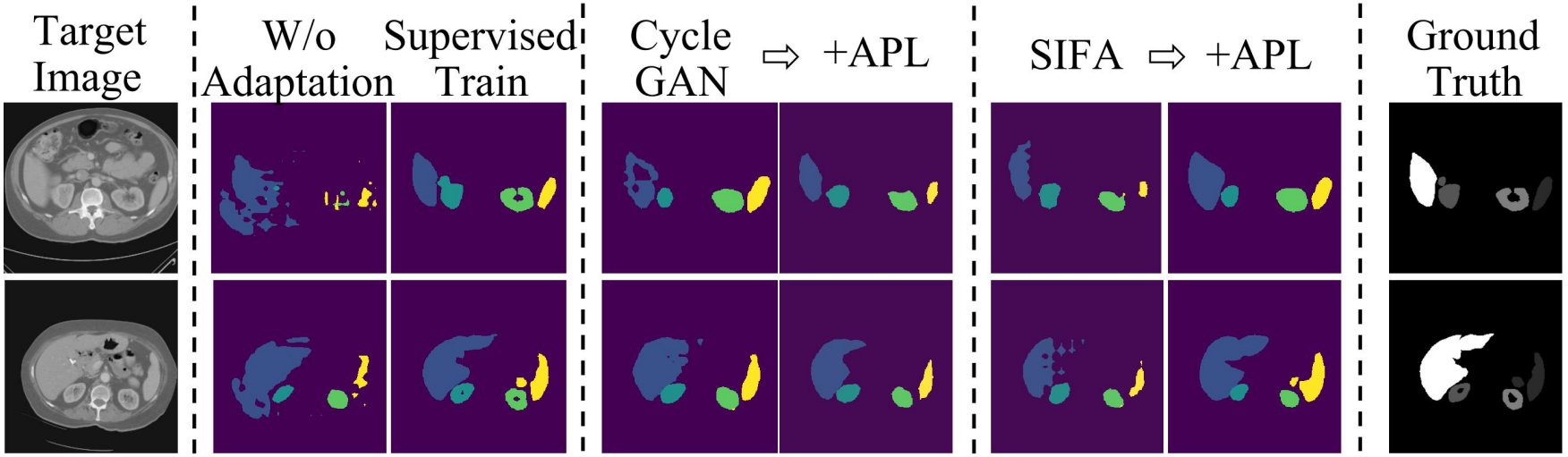
Comparison with pseudo labeling Methods

- Third, we also compared the different promotions with different pseudo labeling methods.

Pseudo Methods	liver	right kidney	left kidney	spleen	Avg
Threshold-based[8]	81.44	75.18	73.18	79.98	77.59
CBST[9]	89.49	80.6	82.14	84.19	84.11
MRNet[10]	87.77	86.67	81.97	83.1	84.88
APL (L_Con)	88.12	89.7	81.87	81.77	85.37
APL (+L_SSL)	89.56	89.74	83.66	82.31	86.32

Qualitatively measure the domain shift

- Finally, the qualitative segmentation results also showed the challenge of UDA in medical image analysis and the effectiveness of APL.



Summary

- We proposed a novel regularization for adaptive pseudo-label denoising:
  - Combining the dual-classifiers consistency and predictive category-aware confidence.
  - The proposed method was orthogonal and thus can be plug and-play to improve existing UDA methods.
  - The pseudo labels were denoised and then used to refine the adapted model without source domain samples.

- We would like to thank:
  - National Key Research and Development Program of China (No. 2018YFB0204301)

- We are grateful for corrections and discussions!

Adaptive Pseudo Labeling For Source-Free Domain Adaptation In Medical Image Segmentation

Thank You!