ADAPTIVE PSEUDO LABELING FOR SOURCE-FREE DOMAIN ADAPTATION IN MEDICAL IMAGE SEGMENTATION

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

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

ABSTRACT
Domain adaptation is common but challenging in signal processing tasks due to the intrinsic discrepancy, especially in difficult-to-label medical image segmentation application scenarios. Pseudo labeling methods are widely utilized to compensate for the scarcity of annotation. However, most existing methods set the fixed thresholds to select highly-confident predictions as pseudo labels, inevitably generating false labels with noise. In this paper, we combine the dual-classifiers consistency and predictive category-aware confidence to form a novel regularization for pseudo-label denoising. The dual-classifiers consistency helps promote the robustness of pseudo labels. Meanwhile, category-aware confidence is utilized as adaptive pixel-wise weights, avoiding the need for handcrafted thresholds. The adapted model is refined by the rectified pseudo labels without source domain samples. The proposed method is model-independent and thus can be plug-and-play to improve existing UDA methods. We validated it on the cross-modality medical image segmentation and obtained more competitive results.

Index Terms—Unsupervised domain adaptation, Source-free, Adaptive pseudo labeling, Medical image segmentation.

1. INTRODUCTION
Deep learning-based methods have brought remarkable advances in the signal processing field, such as image analysis [1, 2, 3, 4]. Supervised learning is one of the most accomplished research, relying on massive annotated data to provide direct supervision for feature extraction and reconstruction. This mechanism is beneficial to promote performance but limits the generalization ability of the networks at the same time. Intuitively, the well-trained networks are prone to suffer from performance degradation when transferred from source domain to target domain. This phenomenon is caused by the feature distribution gap between different domains, which is called domain shift. Domain shift exists in cross-modality image analysis tasks, such as learning the domain-shared structural features from Magnetic Resonance Imaging dataset (MRI) and then segmenting multi-organs from Computed Tomography (CT). When applying the supervised training method to domain generalization, it is challenging to collect sufficient pixel-wise annotations due to privacy protection or unaffordable expense. Meanwhile, it is hard to utilize conventional methods to get promising performance without any target domain labels [5]. In this case, unsupervised domain adaptation (UDA) [6] becomes a prefering choice.

Related work: Most UDA methods try to align domains by minimizing the discrepancy between source domain and target domain in the feature space, where annotations of target samples are absent for training. In collaboration with the above process, pseudo labeling is the self-training method to generate pseudo labels based on the learned domain-invariant representation. Then the pseudo labels are utilized as annotations to fine-tune the adapted model in return. Lee et al. [7] drew on the proposed above idea and leveraged the pseudo labels of unlabeled samples to semi-supervised train the deep neural networks. Zou et al. [8] proposed the CBST and first introduced the idea of pseudo labeling into domain adaptation. Through iterative optimization, the knowledge extracted from source domain is distilled to target domain and achieve domain adaptation. Zou et al. [9] adapted the pseudo labeling method to the semantic segmentation and obtained soft confident results by regularization. Zheng et al. [10] proposed the uncertainty estimation method to calculate the variance between predictions and extract domain-shared features.

Most threshold-based methods generated confident pseudo labels through handcrafted thresholds. However, they ignored that there was still noise inside the prediction. Both noise and semantic information in the prediction were treated equally, which compromised the subsequent training. Meanwhile, the optimal handcrafted thresholds are different for different tasks. Other methods still need to utilize source domain samples for regularization. So it was prone to overfit the source domain-specific features during pseudo labeling. Besides, previous works failed to take advantage of the complementary relationship between confidence and consistency.

This paper focuses on the unsupervised source-free domain adaptation and proposes an adaptive pseudo labeling method for cross-modality medical images segmentation. Specifically, we combine the dual-classifiers consistency and predictive category-aware confidence to form a novel reg-
An overview of our proposed method. Generated from the pre-adapted model, pseudo labels are then denoised with the dual-classifiers consistency. Category-aware confidence is utilized to refine the target model with rectified pseudo labels.

Fig. 1. An overview of our proposed method. Generated from the pre-adapted model, pseudo labels are then denoised with the dual-classifiers consistency. Category-aware confidence is utilized to refine the target model with rectified pseudo labels.

2. METHODOLOGY

2.1. Definition and Motivation

In the unsupervised domain adaptation, we have access to source domain $X_S$ with annotation $Y_S$, where image $x_s \in X_S$ and label $y_s \in Y_S$. Besides, we also have access to target domain $X_T$ without annotation, where image $x_t \in X_T$. It is vital that both $X_S$ and $Y_S$ share the same implicit semantic feature space, where two domains describe the same object with different data distributions. The goal of UDA is to construct projection function to learn the domain-invariant representation with the assistance of $\{X_S, Y_S, X_T\}$. Based on UDA, source-free domain adaptation is to distill the learned representation to target domain without source samples.

After finishing domain adaptation, the adapted model is utilized to predict the pixel-wise classification results of target samples as $\hat{Y}_T$. Existing works mostly calculated the prediction probability and then filtered the pixels with the handcrafted max probability threshold or uncertainty regularization. However, the handcrafted threshold-based methods still require manually design thresholds to separate high-confident objects, which were not an automatic learning process and neglected confident pixels. Meanwhile, 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. These defects greatly limited the robustness of the methods.

For the above reasons, we investigate two critical properties for the generated pseudo labels, i.e., consistency and confidence. **Consistency** is measured by calculating the variance of predictions from different branches and then integrate them for adaptive pseudo labels rectification; (2) The proposed APL method is model-independent and thus can be plug-and-play to promote existing UDA methods without any extra parameters or source domain samples; (3) The proposed APL method is validated on the cross-modality (MRI $\Rightarrow$ CT) medical image segmentation and achieved SOTA performance.
ance between C and D is large, and the maximum values of C and D are also large, which means these pseudo labels are confident but inconsistent. The above two comparisons denote that there is massive noise in the pseudo labels. Directly using these noisy labels as accurate annotations to optimize the model would compromise the subsequent training.

To alleviate the impact of noise, we propose a full-automatic adaptive pseudo-labeling method. Based on the rectified labels, the dynamic pixel-wise weights replace manually designed thresholds. Following most research in UDA, the modified DeepLabv2 with dual-branch classifiers are selected as the baseline model. After obtaining the adapted model from the existing UDA methods, we generated confident pseudo labels with low uncertainty to refine the trained model. The proposed method is orthogonal to the existing UDA methods and yields competitive performance of semantic segmentation.

2.2. Dual-Classifiers Consistency Denoising

Dual-classifiers in the baseline model are composed of decoders with two branch paths, including the main classifier $D_m(\cdot)$ and the auxiliary classifier $D_a(\cdot)$. The process of feature extraction through the encoder is denoted as $E(\cdot)$. Referring to the BCDM proposed by [11], we redesign the loss function to measure the pixel-wise consistency between the prediction results of the $D_m(E(\cdot))$ and $D_a(E(\cdot))$. Both of dual-predictions are matrixes with size $C \times 1$. Firstly, we define the weighted Dual-Classifiers Prediction Relevance $\mathcal{R}$ based on the pixel-wise softmax probabilities:

$$\mathcal{R}_{x_t} = W_{x_t} \times D_m(E(x_t)) \times D_a(E(x_t))^T. \quad (1)$$

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

Obviously, $\mathcal{R}$ is a $C \times C$ matrix, which is used to evaluate the relevance across different classes between the dual-classifiers predictions. $\mathcal{R}_{i,j}(i, j \in \{1, 2, 3, ..., C\})$ is the element in the $i$-th row and $j$-th column of $\mathcal{R}$. The values of the diagonal element $\mathcal{R}_{i,i}(i \in \{1, 2, 3, ..., C\})$ 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. With the help of Relevance $\mathcal{R}$, the pseudo labels are denoised and consistent by maximizing the sum of the diagonal elements. Finally, we define the Dual-classifiers Consistency loss $\mathcal{L}_{Con}$ as the follows:

$$\min_{\theta_{D_m}, \theta_{D_a}} \mathcal{L}_{Con} = \sum_{x_t} \sum_{i,j=1}^{C} [\mathcal{R}_{i,j} - \mathcal{R}_{i,i}], \quad (2)$$

2.3. Category-aware Confidence Denoising

The whole training and optimizing process of our framework is described in Algorithm 1. 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

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

And then, the Category-aware Confidence is calculated and formulated as follows.

$$\omega = |\delta_1(\hat{y}_t) - \delta_2(\hat{y}_t)|, \quad (3)$$

where $\omega$ is the pixel-wise measured confidence, $\delta_1(\hat{y}_t)$ denotes the largest category probability among the pseudo label $\hat{y}_t$. In the same way, $\delta_2(\hat{y}_t)$ is the second largest category probability. $\ell_1$ distance is used to measure the discrepancy between above two probability distributions.

Through the proposed adaptive pseudo labeling method, we automatically obtain the confident labels with high consistency. Based on the denoised labels, we directly utilize them as the original target domain’s annotation and conduct Semi-Supervised Learning to refine the adapted model.

$$\min \mathcal{L}_{SSL} = \sum_{x_t \in X_T, \hat{y}_t \in Y_T} \omega \cdot [-\hat{y}_t \log D(E(x_t))], \quad (4)$$

3. EXPERIMENTS AND RESULTS

To evaluate the effectiveness and superiority of the proposed adaptive pseudo labeling method (APL), extensive experiments are performed on four different medical image segmentation tasks of diverse modalities (MRI ⇒ CT).

3.1. Experimental settings

As for the experimental dataset, the Combined Healthy Abdominal Organ Segmentation (CHAOSS) and Multi-Atlas Labeling Beyond the Cranial Vault (MALBCV) are chosen for cross-modality medical image segmentation. These datasets

---

Table 1. Experimental results of the promoted different domain adaptation methods with the proposed pseudo-labeling method for MRI ⇒ CT abdominal organs segmentation.

<table>
<thead>
<tr>
<th>UDA Methods (Mean)</th>
<th>Dice (%)</th>
<th>LV</th>
<th>RK</th>
<th>LK</th>
<th>SP</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>W/o Adaptation</td>
<td></td>
<td>73.10</td>
<td>47.30</td>
<td>57.30</td>
<td>55.10</td>
<td>58.20</td>
</tr>
<tr>
<td>Supervised Train</td>
<td>92.80</td>
<td>86.40</td>
<td>87.40</td>
<td>88.20</td>
<td>88.70</td>
<td></td>
</tr>
<tr>
<td>SynSegNet [12]</td>
<td>85.00</td>
<td>82.10</td>
<td>72.70</td>
<td>81.00</td>
<td>80.20</td>
<td></td>
</tr>
<tr>
<td>SynSegNet +APL</td>
<td>85.19</td>
<td>81.26</td>
<td>79.24</td>
<td>84.41</td>
<td>82.53</td>
<td></td>
</tr>
<tr>
<td>AdaOutput [13]</td>
<td>85.40</td>
<td>79.70</td>
<td>79.70</td>
<td>81.70</td>
<td>81.60</td>
<td></td>
</tr>
<tr>
<td>AdaOutput +APL</td>
<td>87.65</td>
<td>83.85</td>
<td>83.88</td>
<td>80.77</td>
<td>84.04</td>
<td></td>
</tr>
<tr>
<td>CycleGAN +APL</td>
<td>83.40</td>
<td>79.30</td>
<td>79.40</td>
<td>77.30</td>
<td>79.90</td>
<td></td>
</tr>
<tr>
<td>CyCADA +APL</td>
<td>86.79</td>
<td>78.54</td>
<td>79.24</td>
<td>84.62</td>
<td>82.30</td>
<td></td>
</tr>
<tr>
<td>CyCADA +APL</td>
<td>85.40</td>
<td>78.60</td>
<td>80.30</td>
<td>76.90</td>
<td>80.10</td>
<td></td>
</tr>
<tr>
<td>ADVENT [16]</td>
<td>89.06</td>
<td>81.43</td>
<td>83.30</td>
<td>81.33</td>
<td>83.78</td>
<td></td>
</tr>
<tr>
<td>ADVENT +APL</td>
<td>89.28</td>
<td>77.05</td>
<td>81.37</td>
<td>83.45</td>
<td>82.79</td>
<td></td>
</tr>
<tr>
<td>SIFA-v1 +APL</td>
<td>90.37</td>
<td>85.10</td>
<td>81.48</td>
<td>84.82</td>
<td>85.44</td>
<td></td>
</tr>
<tr>
<td>SIFA-v2 +APL</td>
<td>87.90</td>
<td>83.70</td>
<td>80.10</td>
<td>80.50</td>
<td>83.10</td>
<td></td>
</tr>
<tr>
<td>SIFA-v2 +APL</td>
<td>89.56</td>
<td>89.74</td>
<td>83.66</td>
<td>82.31</td>
<td>86.32</td>
<td></td>
</tr>
<tr>
<td>SIFA-v2 +APL</td>
<td>88.00</td>
<td>83.30</td>
<td>80.90</td>
<td>82.60</td>
<td>83.70</td>
<td></td>
</tr>
<tr>
<td>SIFA-v2 +APL</td>
<td>89.56</td>
<td>87.88</td>
<td>89.19</td>
<td>82.80</td>
<td>87.11</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of quantitative results between our APL method and other SOTA pseudo-label methods on cross-modality abdominal organ segmentation. Dice (%) of four organs and their average are compared here.

<table>
<thead>
<tr>
<th>Pseudo Methods</th>
<th>Dice (%)</th>
<th>LV</th>
<th>RK</th>
<th>LK</th>
<th>SP</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold [7]</td>
<td></td>
<td>81.44</td>
<td>75.18</td>
<td>73.18</td>
<td>79.98</td>
<td>77.59</td>
</tr>
<tr>
<td>CBST [8]</td>
<td>89.49</td>
<td>80.60</td>
<td>82.14</td>
<td>84.19</td>
<td>84.11</td>
<td></td>
</tr>
<tr>
<td>MRNet [10]</td>
<td>87.77</td>
<td>86.67</td>
<td>81.97</td>
<td>83.10</td>
<td>84.88</td>
<td></td>
</tr>
<tr>
<td>APL (Lcon)</td>
<td>88.12</td>
<td>89.70</td>
<td>81.87</td>
<td>81.77</td>
<td>85.37</td>
<td></td>
</tr>
<tr>
<td>APL (+LSSL)</td>
<td>89.56</td>
<td>89.74</td>
<td>83.66</td>
<td>82.31</td>
<td>86.32</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Comparison of segmentation results produced by different methods for abdominal organs segmentation.

are derived from the public MICCAI and ISBI challenges, containing multi-abdominal organs described in different modalities, including 30 CT volumes and 20 MRI volumes.

As for the preprocessing method, the volume data is sliced as 2D images through transverse view and normalized with the size of 256x256. The segmentation objects consist of four abdominal organs, including liver (LV), right kidney (RK), left kidney (LK), and spleen (SP). Among the preprocessed data, the labels of four organs are obtained through pixel-level annotation by experienced specialists. As for the evaluation metric, the Dice similarity coefficient [19, 20] is utilized to compare the performance discrepancy of methods by calculating the overlap between predicted result and ground truth. Higher Dice values represent better performance.

3.2. Promotion based on adaptive pseudo labeling method

Firstly, we quantitatively measured the domain shift by calculating performance dependency in the abdominal organs segmentation. The bottom bound was to directly transfer the well-trained model from source domain to target domain without any domain adaptation technique. The top bound was to use the labels of target samples to train the target model based on the supervised train. Both of their performance were reported in the Table 1. We can find the enormous performance gap between top bound and bottom bound, which also indicated the severe domain shift between cross-modality images (MRI ⇒ CT) and the challenge of unsupervised source-free domain-adaptive medical image segmentation.

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. The quantitative results were reported in the Table 1, where the results of the above SOTA methods all referred from the paper [18]. Compared with the original adapted models based on the UDA methods, the introduced APL method made remarkable improvements by refining these models with rectified pseudo labels. For example, SIFA [17] increased the average Dice from 83.10% to 86.32% after introduced the APL method.

Third, we also compared the different promotions with different pseudo labeling methods. Adapted SIFA-v1 [17] was used as the basic model. The results were reported in Table 2. Our APL consistently outperformed the state-of-the-art methods as well as existing handcrafted-threshold strategies.

Finally, the qualitative segmentation results in Fig.2 also showed the challenge of UDA in medical image analysis. Our method can successfully locate the four organs and generate semantically meaningful masks. All the above qualitative and quantitative results validated the effectiveness of our method.

4. CONCLUSION

In this paper, we proposed a novel regularization for adaptive pseudo-label denoising, combining the dual-classifiers consistency and predictive category-aware confidence. Besides, 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. The experimental results demonstrated that the proposed method was superior to other methods in the cross-modality medical image segmentation.
5. REFERENCES


