STAMP: A Self-training Student-Teacher Augmentation-Driven Meta Pseudo-Labeling Framework for 3D Cardiac MRI Image Segmentation

S. M. Kamrul Hasan[1(✉)] and Cristian Linte[2]

[1] Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, USA
sh3190@rit.edu

[2] Biomedical Engineering, Rochester Institute of Technology, Rochester, NY, USA

Abstract. Medical image segmentation has significantly benefitted thanks to deep learning architectures. Furthermore, semi-supervised learning (SSL) has led to a significant improvement in overall model performance by leveraging abundant unlabeled data. Nevertheless, one shortcoming of pseudo-labeled based semi-supervised learning is pseudo-labeling bias, whose mitigation is the focus of this work. Here we propose a simple, yet effective SSL framework for image segmentation-*STAMP* (*Student-Teacher A*ugmentation-driven consistency regularization via *M*eta *P*seudo-Labeling). The proposed method uses self-training (through meta pseudo-labeling) in concert with a Teacher network that instructs the Student network by generating pseudo-labels given unlabeled input data. Unlike pseudo-labeling methods, for which the Teacher network remains unchanged, meta pseudo-labeling methods allow the Teacher network to constantly adapt in response to the performance of the Student network on the labeled dataset, hence enabling the Teacher to identify more effective pseudo-labels to instruct the Student. Moreover, to improve generalization and reduce error rate, we apply both strong and weak *data augmentation* policies, to ensure the segmentor outputs a consistent probability distribution regardless of the augmentation level. Our extensive experimentation with varied quantities of labeled data in the training sets demonstrates the effectiveness of our model in segmenting the left atrial cavity from Gadolinium-enhanced magnetic resonance (GE-MR) images. By exploiting unlabeled data with weak and strong augmentation effectively, our proposed model yielded a statistically significant 2.6% improvement ($p < 0.001$) in Dice and a 4.4% improvement ($p < 0.001$) in Jaccard over other state-of-the-art SSL methods using only 10% labeled data for training.

Research reported in this publication was supported by the National Institute of General Medical Sciences Award No. R35GM128877 of the National Institutes of Health, and the Office of Advanced Cyber infrastructure Award No. 1808530 of the National Science Foundation.

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 371–386, 2022.
https://doi.org/10.1007/978-3-031-12053-4_28

Keywords: Meta pseudo-label · Cardiac MRI segmentation · Weak and strong augment · Confidence threshold · Student-teacher model

1 Introduction

While deep learning has shown potential for improved performance across a wide variety of medical computer vision tasks, including segmentation [4,5], registration [2], and motion estimation [25], many of these successes are achieved at the cost of a large pool of labeled datasets. Obtaining labeled images, on the other hand, requires substantial domain expertise and manual labor, making large-scale deep learning models challenging to implement in clinical settings. Moreover, when the annotation of medical images requires the assistance of clinical experts, the cost becomes unaffordable. Hence, this ineffectiveness in the low-data domain, in turn, hampers the clinical adoption and use of many medical image segmentation models. Therefore, instead of attempting to improve high-data regime segmentation, this work focuses on data-efficient segmentation training that only uses a few pixel-labeled data and takes advantage of the wide availability of unlabeled data to improve segmentation performance, with the goal of closing the performance gap with supervised models trained with fully pixel-labeled data.

Our work is motivated by the recent progress in image segmentation using semi-supervised learning (SSL), which has shown good results with limited labeled data and large amounts of unlabeled data. Recent research has yielded a variety of semi-supervised learning techniques. Successful examples include MeanTeacher [20], MixMatch [3], and FixMatch [19]. One outstanding key feature of most SSL frameworks is consistency regularization, which encourages the model to produce the same output distribution when its inputs are perturbed [7,16]. As such, pseudo-labeling or self-training is also utilized in conjunction with semi-supervised segmentation to incorporate the model's own predictions into the training [1,11]. As such, to increase training data, models incorporate pseudo-labels of the unlabeled images obtained from the segmentation model trained on the labeled images.

To execute a task, semi-supervised learning (SSL) uses a small number of labeled examples along with unlabeled samples. Most methods follow one or combinations of directions, such as consistency regularization [18,19] or pseudo-labeling [9,11]. Existing methods use conventional data augmentation [10,20] to provide alternative transformations of semantically identical images, or they blend input data to create enhanced training data and labels [8,23]. Liu *et al.* [13] revisit the Semi-Supervised Object Detection and identify the pseudo-labeling bias issue in SS-OD. However, they updated the Teacher network using a non-gradient exponential moving average (EMA), which concentrates on weighting the Student's parameters at each stage of the training process, without explicitly evaluating parameter quality. Sohn *et al.* introduce FixMatch [19], which matches the prediction of the strongly-augmented unlabeled data to the pseudo label of the weakly-augmented counterpart when the model confidence on the

weakly-augmented counterpart is high. In contrast to these approaches, here we redesign the pseudo label as well as data augmentations for semantic segmentation utilizing both consistency regularization, as well as pseudo labeling.

A self-training based approach was used by Bai *et al.* [1] for cardiac MR image segmentation. They use an initial model trained on labeled data to predict the labels on unlabeled data, so that these labels, although less accurate, can be used for training an updated, more powerful model. Recent approaches involve integrating uncertainty map into a mean-Teacher framework to guide the Student network [22] for left atrium segmentation. Zeng *et al.* [24] propose a Student-Teacher framework for semi-supervised left atrium segmentation. However, they haven't applied any data augmentation and thus omit the idea that a segmentor should output the same probability distribution for an unlabeled pixel even after it has been augmented.

Nevertheless, pseudo-labeling techniques, despite their benefit, suffer from one major flaw: if the pseudo-labels are erroneous, the Student network will learn from inaccurate data, much like the analogy of a Student's performance (i.e., the accuracy of the segmentation labels output by a model) not being able to significantly exceed the Teacher's performance (i.e., the accuracy of the pseudo-labels used for training the model). This flaw is also known as the problem of confirmation bias in pseudo-labeling. To this extent, this paper investigates pseudo-labeling for semi-supervised deep learning from network predictions and shows that in contrast to previous attempts at pseudo-labeling [15,24], simple modifications to correct confirmation bias results in state-of-the-art performance.

To address these issues, we propose a three-stage semi-supervised framework - ***STAMP: Student-Teacher Augmentation-Driven Meta Pseudo-Labeling***, inspired by the framework in Noisy-Student [21], a method of training

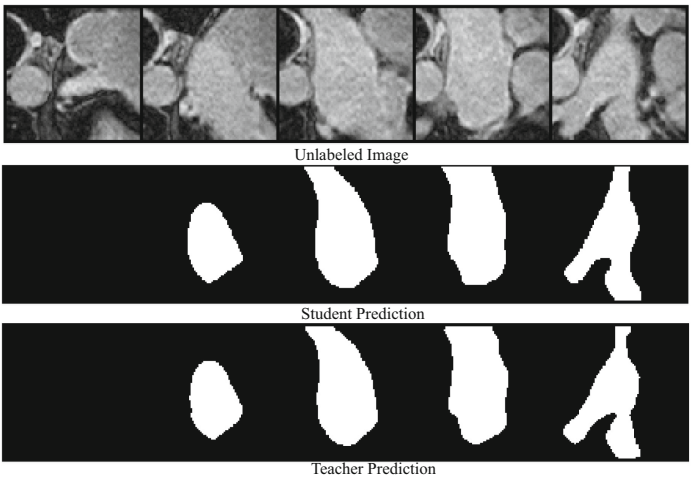

Fig. 1. STAMP model applied to the left atrium dataset, where a large amount of unlabeled data is available. Both the Student and Teacher predictions are shown during a random training iteration.

a Student and a slowly progressing Teacher (Fig. 1) in a mutually advantageous manner. In the first stage, we train a fully convolutional network (FCN) using all labeled data until convergence. In the second stage, the weak data augmentations are applied to each unlabeled image where the Teacher model is trained with unlabeled data and the Student learns from a minibatch of pseudo-labeled data generated by the Teacher. The prediction of strongly-augmented data is then optimized to match its corresponding pseudo-labels with the labeled data pre-trained in the first stage. Later on, the Student progressively updates the Teacher using the response signal in the third stage. Unlike the non-gradient EMA [10] method, this reward signal is utilized to motivate the Teacher during the Student's learning process through a gradient descent algorithm. We evaluate our approach using the Left Atrial Segmentation Challenge dataset by comparing our results to those of existing SSL methods. STAMP achieves a 2.6 fold mean improvement over the state-of-the-art RLSSS [24] method.

Our proposed method presents several key contributions which are summarized as follows: (1) *STAMP* presents simple and effective strategy for dealing with the pseudo-labeling bias problem by adopting a *threshold* where pixels with a confidence score higher than 0.5 will be used as pseudo labels, while the remaining are treated as ignored regions. Additionally, since a large pool of labeled data is not available, the proposed method inherently mitigates the over-fitting problem; (2) The different strong and weak *data augmentation* policies improve the generalization performance and reduce the error rate significantly. Our observation shows that when replacing weak augmentation with no augmentation, the model overfits the predicted unlabeled labels; (3) The use of pseudo-labels enables a *gradient descent* response loop from the Student network to the Teacher network that improves the teaching of the Teacher network and minimizes the prediction bias; and (4) Extensive experimental studies on the MICCAI STACOM 2018 *Atrial segmentation* challenge dataset and comparative analyses are conducted to validate the effectiveness of this method at not only the low-data regime, but also the high-data regime.

2 Methodology

2.1 STAMP Model Framework

2.1.1 Segmentation Model Formulation

We define the semi-supervised image segmentation problem in a semi-supervised setting as follows: given an (unknown) data distribution $p(x, y)$ over images and segmentation masks, we have a source domain having a training set, $\mathcal{D}_{\mathcal{L}} = \{(x_i^l, y_i^l)\}_{i=1}^{n_l}$ with n_l labaled examples and $\mathcal{D}_{\mathcal{UL}} = \{(x_j^{ul})\}_{j=1}^{n_{ul}}$ with n_{ul} unlabaled examples which are sampled i.i.d. from $p(x, y)$ and $p(x)$ distribution and $n_l \ll n_{ul}$, where x_i^l is the i-th labeled image with spatial dimensions $H \times W$, $y_i^l \in \{0, 1\}^{C \times H \times W}$ is its corresponding pixelwise label map with C as the number of categories, and x_j^{ul} is the j-th unlabeled image. Empirically, we want to minimize the target risk $\phi_t(\theta^S, \theta^T) = min_{\theta^S, \theta^T} \quad \mathcal{L}_{\mathcal{L}}(\mathcal{D}_{\mathcal{L}}, (\theta^S, \theta^T)) + \gamma \mathcal{L}_{\mathcal{UL}}(\mathcal{D}_{\mathcal{UL}}, (\theta^S, \theta^T))$,

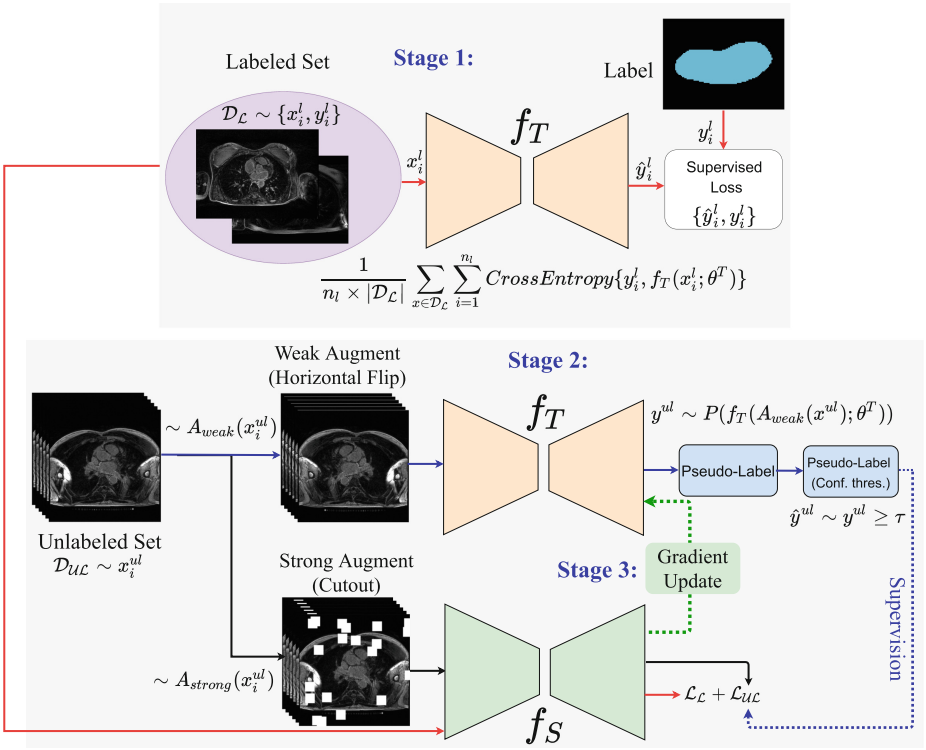

Fig. 2. Schematic of *STAMP* model: The Teacher model is trained using all labeled data until convergence. Weak data augmentations are applied to each unlabeled image, such that the Teacher model is trained with unlabeled data and the Student learns from a mini-batch of pseudo-labeled data generated by the Teacher. In turn, the Teacher's parameters θ_T are updated based on the response signal from the Student's parameters θ_S via gradient-descent in the later stage.

where $\mathcal{L}_\mathcal{L}$ is the supervised loss for segmentation, $\mathcal{L}_\mathcal{UL}$ is unsupervised loss defined on unlabeled images and θ^S, θ^T denotes the learnable parameters of the overall network.

2.1.2 Model Architecture and Components

We propose **STAMP** – a simple yet effective **S**tudent-**T**eacher SSL framework for image segmentation based on **A**ugmentation driven Consistency regularization and Self-Training (through **M**eta **P**seudo-labeling), as illustrated in Fig. 2. The overall model entails three stages of training, where we train a Teacher model using all available labeled data in the first stage as a pre-trained initializer, while in the second stage, we train STAMP using both labeled and unlabeled data. We manage the quality of pseudo labels constituted of segmentation masks using a high confidence-based threshold value inspired by FixMatch [19]. The training steps for STAMP are summarized in the subsequent sections.

(a) Training a Teacher Model: It is critical to start with an appropriate initialization for both the Student and Teacher models because we'll be relying on the Teacher to create pseudo-labels to subsequently train the Student. Hence, we first apply the supervised loss $\mathcal{L}_{\mathcal{L}}$ to improve our model using the existing supervised data. For a labeled set $\mathcal{D}_{\mathcal{L}} = \{(x_i^l, y_i^l)\}_{i=1}^{n_l}$, the segmentation network is trained in a traditional supervised manner which minimizes the cross-entropy (CE) loss, $\mathcal{L}_{\mathcal{L}} = \frac{1}{n_l \times |\mathcal{D}_{\mathcal{L}}|} \sum_{x \in \mathcal{D}_{\mathcal{L}}} \sum_{i=1}^{n_l} CrossEntropy\{y_i^l, f_T(x_i^l; \theta^T)\}$, where the definitions of parameters are defined in Problem Description section.

(b) Generating Pseudo-Labels: STAMP assigns each unlabeled example an artificial label, which is subsequently employed in a standard cross-entropy loss to train the Student model. We initially compute the model's predicted distribution using a *weakly-augmented* (e.g. horizontal flip) version of a given unlabeled image x_j^{ul} in an unlabeled set $\mathcal{D}_{\mathcal{UL}}$ to obtain an artificial label, $y^{ul} \sim P(f_T(A_{weak}(x^{ul}); \theta^T))$. To avoid the cumulatively detrimental effect of noisy pseudo-labels (i.e., confirmation bias), we first set a confidence threshold τ of predicted masks to filter low-confidence predicted masks, which are more likely to be false-positive samples. Then, the final pseudo-labels are obtained by selecting the pixels having the maximum predicted probability of the corresponding class, $\hat{y^{ul}} = (argmax(P(f_T(A_{weak}(x^{ul})); \theta^T)) \geq \tau)$, where A_{weak} denotes the weak-augmentation operation.

(c) Student Learning from Pseudo-Labels: In this stage, the Student model $f_S(., \theta^S)$ is trained with the pseudo-labels generated from the Teacher model, where we use both the labeled and unlabeled datasets $\mathcal{D}_{\mathcal{L}}$, $\mathcal{D}_{\mathcal{UL}}$. We enforce the cross-entropy loss against the Student model's output for the *strong-augmentation* of the unlabeled images having the idea that the Student model would output the same probability distribution for an unlabeled pixel even after it has been augmented. Additionally, we utilize a consistency regularizer function to enforce consistency between the generated pseudo masks and the masks predicted by the Student model itself (Eq'n 1).

$$\underbrace{\frac{1}{n_{ul} \times |\mathcal{D}_{\mathcal{UL}}|} \sum_{x \in \mathcal{D}_{\mathcal{UL}}} \sum_{j=1}^{n_{ul}} CrossEntropy\{\hat{y_i^{ul}}, f_S(A_{strong}(x_j^{ul}); \theta^S)\} + \sum_{x_i \in \mathcal{D}} ||(\hat{y^{ul}}) - (f_S(A_{strong}(x_j^{ul}); \theta^S))||^2}_{\textbf{Regularizer}} \tag{1}$$

where A_{strong} denotes the strong-augmentation (Cutout, Gaussian blur, ShiftScaleRotate) operation. Since the Student parameters always depend on the Teacher parameters via the pseudo labels, we need to compute the Jacobian, as shown in Eq'n (2) (Algorithm 1).

Algorithm 1. STAMP's main learning algorithm

Input:
Training set of labeled data x^l, $y^l \epsilon \mathcal{D}_{\mathcal{L}}$, and unlabeled data $x^{ul} \epsilon \mathcal{D}_{\mathcal{UL}}$
Require: Learned parameters: (θ^T, θ^S), number of pre-train epoch, number of main-train epoch, confidence threshold, τ
for *each epoch* **do**
  if $epoch < main_{train}$ **then**
    Sample mini-batch from $x_i^l; x_1^l, \ldots, x_{n_l}^l$;
    $\theta^T \leftarrow \theta^T + \gamma \frac{\partial L_{sup}}{\partial \theta^T}$ {Train the Teacher network with all the labeled data}
  else
    Teacher UPDATE STAGE:
    Sample mini-batch from $x_i^l; x_1^l, \ldots, x_{n_l}^l$; and $x_j^{ul}; x_1^{ul}, \ldots, x_{n_{ul}}^{ul}$;
    Apply weak data augmentation to $x^{ul}, x^{ul} = A_{weak}(x^{ul})$ to train the Teacher model
    Apply strong data augmentation to x^{ul}, $x^{ul} = A_{strong}(x^{ul})$ to train the Student model
    Sample a pseudo label $y^{ul} \sim P(f_T(A_{weak}(x^{ul}); \theta^T))$
    Use a confidence threshold, τ
    if $P(f_T(A_{weak}(x^{ul}); \theta^T)) \geq \tau$ **then**
      pseudo-mask, $\hat{y^{ul}} = argmax(y^{ul})$
    end if
    Update the Student using the pseudo label $\hat{y^{ul}}$:

$$\theta^S_{(t+1)} = \theta^S_{(t)} - \eta S \ \nabla_{\theta^S} \ CE(\hat{y^{ul}}, f_S((A_{weak}(x^{ul}); \theta^S))|_{\theta^S = \theta^S_{(t)}} \tag{2}$$

    Compute the Teacher's response coefficient

$$h = \eta S. \ \Big((\nabla_{\theta'^S} \ CE(y^l, f_S(x^l; \theta^S_{(t+1)})))^\top . \\ \nabla_{\theta^S} CE(\hat{y^{ul}}, f_S(A_{weak}(x^{ul}); \theta^S)) \Big) \tag{3}$$

    Compute the Teacher's gradient from the Student's response signal:

$$g^T_{(t)} = h. \ \nabla_{\theta^T} CE(\hat{y^{ul}}, f_T(\mathcal{A}(x^{ul}); \theta^T))|_{\theta^T = \theta^T_{(t)}} \tag{4}$$

    Compute the Teacher's gradient on labeled data:

$$g^{T,Sup}_{(t)} = \nabla_{\theta^T} CE(y^l, f_T(x^l; \theta^T)) \tag{5}$$

    Update the Teacher:

$$\theta^T_{(t+1)} = \theta^T_{(t)} - \eta T. \Big(g^T_{(t)} + g^{T,Sup}_{(t)} \Big) \tag{6}$$

  end if
end for

(d) Updating the Teacher Model: To obtain more stable meta pseudo-labels, we use the response signal from the Student to gradually update the Teacher model. Unlike the non-gradient EMA [10] method, this reward signal is

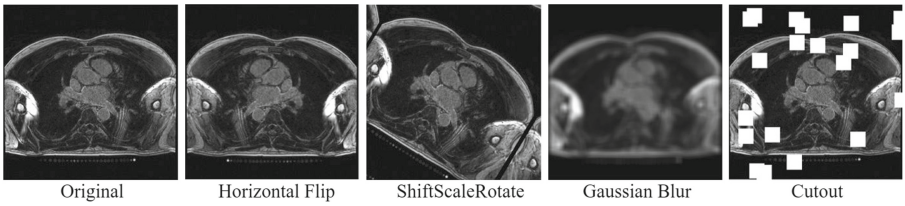

Fig. 3. Visualization of different types of augmentation strategies. Original image, Horizontal Flip, ShiftScaleRotate, Gaussian Blur, and Cutout (left to right).

utilized to motivate the Teacher during the Student's learning process through the gradient descent algorithm as described in [17] (Eq'n 3–6).

2.2 Data Augmentation Strategies

A robust data augmentation is a vital aspect in the success of SSL approaches like MixMatch [3], FixMatch [19] etc. We leverage the Cutout augmentation [6] (strong augmentation) with a rectangle of 50×50 pixels because of its consistent results. We investigate various transformation techniques including Horizontal Flip (weak augmentation), Gaussian Blur, ShiftScaleRotate colorJitter, etc. Each operation has a magnitude that determines the degree of strength augmentation. We visualize transformed images with the aforementioned augmentation strategies in Fig. 3.

2.3 Experiments

Data: The model was trained and tested on the MICCAI STACOM 2018 Atrial Segmentation Challenge datasets featuring 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) and LA segmentation masks, with an isotropic resolution of $0.625 \times 0.625 \times 0.625\,\text{mm}^3$. The dimensions of the MR images vary depending on each patient, however, all MR images contain exactly 88 slices in the z axis. All the images were normalized and resized to $112 \times 112 \times 80$ before feeding them to the models. We split the dataset into 80 scans for training and 20 scans for validation, and apply the same pre-processing methods.

Baselines Architecture: For a fair comparison, we use V-Net [14] as the backbone for both the Teacher and the Student models in our semi-supervised segmentation experiments.

Training: The performance of semi-supervised models trained for image segmentation can significantly be enhanced by the selection of the regularizer, optimizer, and hyperparameters. We implement our method using the *PyTorch* framework and set the batch size to 4. In self-training, a batch of 4 images is composed of 2 labeled images and 2 unlabeled images. Both the Teacher and the Student models are trained for 6000 iterations, with an initial learning rate 0.01, decayed by 0.1 every 2500 iterations. We train the network on varying

proportions of labeled data – 10%, 20%, 30%, 50%, and 100% – while enforcing that $|\mathcal{D}_{\mathcal{L}}| \leq |\mathcal{D}_{\mathcal{UL}}|$. We include an ablation study to elucidate and investigate the effects of the different components and hyperparameters of our model. All experiments were conducted on a workstation equipped with two NVIDIA RTX 2080 Ti GPUs (each 11GB memory). The detailed training procedure is presented in Algorithm 1.

2.4 Evaluation

To evaluate the performance of semantic segmentation of cardiac structures, we use several standard metrics, including Dice score (Dice), Jaccard index, Hausdorff distance (HD), Precision, and Recall. We compare the segmentation results achieved using our proposed *STAMP* architecture with those achieved using five other frameworks: V-Net, MT, UA-MT, SASSNet, and RLSSS.

To justify the choice of these frameworks as benchmarks, here we briefly highlight their features. The UA-MT [22] model is based on the uncertainty-aware mean Teacher framework, in which the Student model learns from meaningful targets over time by leveraging the Teacher model's uncertainty information. The Teacher model not only generates the target outputs, but it also uses Monte Carlo sampling to quantify the uncertainty of each target prediction. When computing the consistency loss, they use the estimated uncertainty to filter out the faulty predictions and keep only the dependable ones (low uncertainty).

Similarly, to take advantage of the unlabeled data and enforce a geometric form constraint on the segmentation output, SASSNet [12] offered a shape-aware semi-supervised segmentation technique. Meanwhile, in semi-supervised image segmentation, self-ensembling approaches, particularly the mean Teacher (MT) model [20], have received a lot of attention. The mean Teacher (MT) structure guarantees consistency of predictions with inputs under varied perturbations between the Student and Teacher models, boosting model performance even more. In RLSSS [24], the Teacher updates its parameters autonomously according to the reciprocal feedback signal of how well Student performs on the labeled set.

3 Results and Discussion

3.1 Image Segmentation Evaluation

We first evaluate our proposed framework on Left Atrium MRI dataset. The quantitative comparison of various approaches in terms of Dice score (Dice), Jaccard index, Hausdorff distance (HD), Precision, and Recall is shown in Table 1. A better segmentation yields a higher Dice, Jaccard, Precision and Recall values and lower values for the other metrics. All semi-supervised approaches that take advantage of un-annotated images enhance segmentation performance significantly when compared to fully-supervised V-Net trained with only 8 (10%) annotated images.

Table 1. Quantitative comparison of left atrium segmentation across several frameworks. Mean (standard deviation) values are reported for $Dice(\%), Jaccard(\%), 95HD(\%), ASD(\%), Precision(\%)$, and $Recall(\%)$ from all networks against our proposed STAMP. The statistical significance of the STAMP results compared to those achieved by the other top performing models, including RLSSS, for 10% and 20% labeled data are represented by $*$ and $**$ for $p-$values 0.1 and 0.001, respectively. The best performance metric is indicated in **bold** text.

METHODS	SCANS USED		METRICS		
	Labeled	Unlabeled	Dice(%) ↑	Jaccard(%)↑	HD95(mm) ↓
V-Net [14]	10%	0	79.98 ±1.88	68.14±2.01	21.12±15.19
MT [20]	10%	90%	83.76±1.03	73.01±1.56	14.56±14.03
UA-MT [22]	10%	90%	84.25±1.61	73.48±1.73	13.84±13.15
SASSNet [12]	10%	90%	87.32±1.39	77.72±1.49	12.56±11.30
RLSSS [24]	10%	90%	88.13±1.68	79.20±1.78	11.59±9.28
STAMP (Proposed)	10%	90%	**90.43±0.75	**82.67±.82	**6.22±4.55
V-Net [14]	20%	0	85.64±1.73	75.40±1.84	16.96±14.37
MT [20]	20%	80%	88.23±1.01	79.29±1.80	10.64±9.32
UA-MT [22]	20%	80%	88.88±0.73	80.20±0.82	8.13±6.78
SASSNet [12]	20%	80%	89.54±0.66	81.24±0.75	8.24±6.58
RLSSS [24]	20%	80%	90.07±0.76	82.03±0.84	**6.67±3.54**
STAMP (Proposed)	20%	80%	*91.90±0.64	**84.38±0.83	7.15±4.74

METHODS	SCANS USED		METRICS		
	Labeled	Unlabeled	ASD(mm)↓	Precision(%) ↑	Recall(%)↑
V-Net [14]	10%	0	5.47±1.92	83.67±1.79	74.55±1.90
MT [20]	10%	90%	4.43±1.08	87.23±1.06	76.31±1.88
UA-MT [22]	10%	90%	3.36±1.58	87.57±1.53	77.85±1.65
SASSNet [12]	10%	90%	2.55±1.86	87.66±1.38	87.22±1.37
RLSSS [24]	10%	90%	2.91±0.59	90.33±1.66	87.08±1.70
STAMP (Proposed)	10%	90%	*1.82±0.40	**90.96±0.74**	**90.30±0.75**
V-Net [14]	20%	0	4.03±1.53	88.78±1.70	83.79±1.51
MT [20]	20%	80%	2.66±1.26	89.89±0.92	87.54±0.66
UA-MT [22]	20%	80%	2.35±1.16	89.57±0.73	88.82±0.72
SASSNet [12]	20%	80%	2.27±0.81	89.86±0.65	90.42±0.66
RLSSS [24]	20%	80%	2.11±4.67	90.16±**0.77**	89.97±**0.76**
STAMP (Proposed)	20%	80%	**2.04±0.34**	**90.92**±0.93	*91.43±0.92

Our proposed model outperformed the fully supervised method according to all metrics, achieving a 90.4% Dice and 82.7% Jaccard scores, which represent a 13% and 21.3% improvement, respectively. Moreover, in comparison to other methods, our proposed framework more efficiently utilized the limited labeled data by employing a Teacher-Student mutual learning strategy, which allowed the Teacher model to update its parameters autonomously and generate more reliable annotations for unlabeled data.

The paired statistical test reported in Table 1 shows that our proposed model significantly improved the segmentation performance compared to the semi-supervised, fully-supervised, models in terms of the Dice, Jaccard, 95% Hausdorff Distance (95HD), average surface distance (ASD), Precision, and Recall. In addition, by effectively exploiting unlabeled data with weak and strong augmentation, our proposed model yielded a statistically significant 2.6%improvement ($p < 0.05$) in Dice and 4.4% Jaccard ($p < 0.05$) over the RLSSS framework, while using *only* 10% labeled data for training.

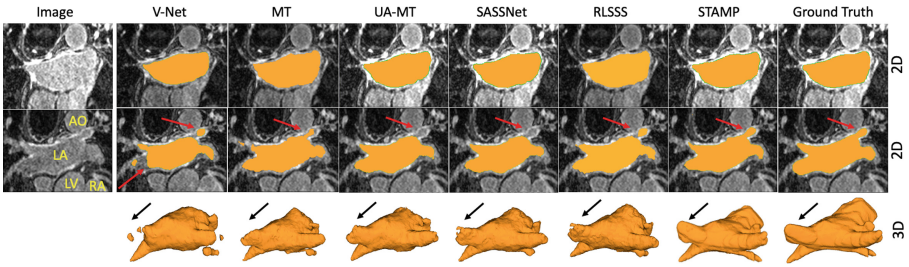

Fig. 4. Qualitative comparison result in 2D as well as 3D of the MICCAI STACOM 2018 Atrial Segmentation challenge dataset yielded by six different frameworks (V-Net, MT, UA-MT, SASSNet, RLSSS, and STAMP). The comparison of segmentation results between the proposed method and five typical deep learning networks indicates that the performance of our proposed network is superior. The black arrows indicate the locations where the segmentation masks yielded by the other networks used as benchmarks fail to correctly capture the aorta (AO) in 3D.

Figure 4 shows the results obtained by V-Net [14], MT [20], UA-MT [22], SASSNet [12], RLSSS [24], our proposed STAMP framework, and the corresponding ground truth on the MICCAI STACOM 2018 Atrial Segmentation Challenge. Figure 4 (bottom row) also shows that all frameworks but STAMP yield segmentation masks that miss portions of the aortic (AO) region (indicated by the red arrows in 2D and black arrows in 3D). On the other hand, the STAMP framework yields a complete segmentation of the left atrium that closely matches the ground truth segmentation mask, preserves more details, and yields fewer false positive results, overall demonstrating the increased efficacy of the proposed learning strategy.

Figure 5(a) shows the best segmentation contours yielded by the STAMP framework (green) and the corresponding ground truth contours (red). We trained our model on varying proportions of labeled data – 10%, 20%, 30%, 50%, and 100% – while enforcing that $|\mathcal{D}_{\mathcal{L}}| \leq |\mathcal{D}_{\mathcal{UL}}|$. Figure 5(b) shows that STAMP accuracy further increases with increasing proportions of labeled data for training. The mean Dice score (%) increases from 90% with only 10% labeled data to 93% with 100% labeled data. This experiment clearly emphasizes the robustness and high performance of STAMP using mostly (90%) unlabeled data, and its *only incremental improvement* with the addition of large quantities of labeled data.

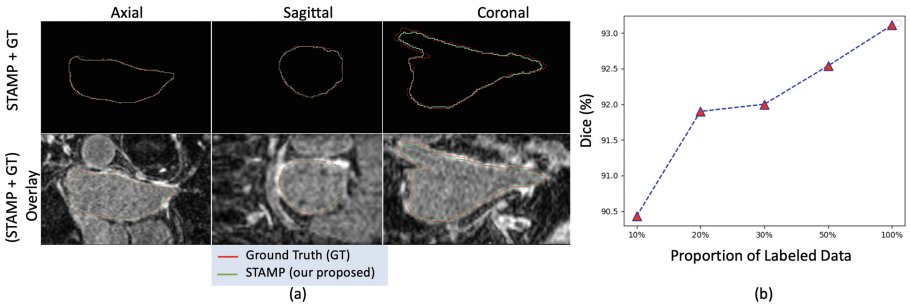

Fig. 5. (a) Axial, coronal and sagittal views of the of the STAMP (green) and ground truth (red) left atrium segmentation contours; (b) robust and high performance (90% Dice score) STAMP segmentation with 10%: 90% labeled: unlabeled data and consistent steady performance increase (up to 93% Dice score) with additional labeled data. (Color figure online)

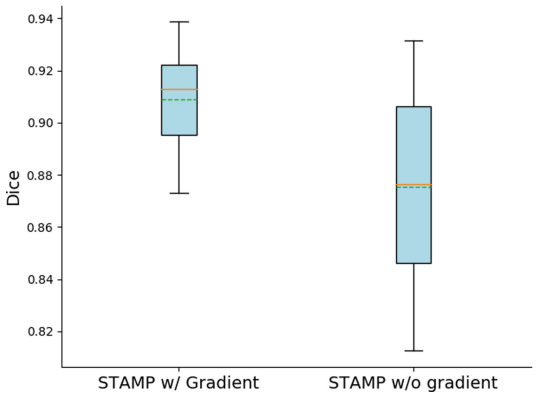

Fig. 6. Ablation study designed to investigate the effect of gradient-based teacher training (GTT) on Dice score for left atrial segmentation using only 20% labeled data with and without GTT.

3.2 Ablation Study

We also conducted ablation studies to demonstrate the effectiveness of incorporating a response signal loop by *gradient descent* step from the Student network to the Teacher network to improve the teaching of the Teacher network and minimize the prediction bias in a semi-supervised setting, as well as study the benefit of different forms of augmentation.

3.2.1 Effect of the Gradient-based Teacher Training

To illustrate the impact of *Gradient-based Teacher training (GTT)*, we compared our model performance with and without GTT. Figure 6 shows that the incorporation of GTT significantly improves segmentation performance, as

quantified by the Dice score. This significant improvement can be explained by the fact that while conventional training (without GTT) often generates imbalanced pseudo-labels, where most pixel category instances in the pseudo-labels vanish, leaving just instances of specific pixel categories, GTT constrains the generation of imbalanced pseudo-labels, leading to improved performance.

3.2.2 Effect of Pre-Training Stage

For both the Student and Teacher models, a proper initialization is critical. Figure 7 shows the effects of using a pre-training stage. We observe that using the *pre-training step*, the model may generate more accurate pseudo-labels early in the training process. As a result, the model can attain lower loss in the training process, as well as better performance once the model converges.

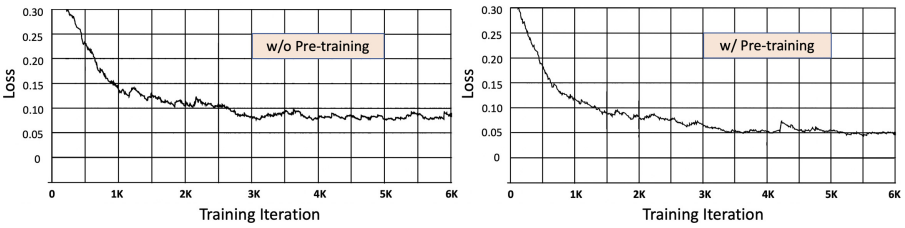

Fig. 7. Experiment conducted on a left atrial image datasets consisting of only 20% labeled data showing the benefits of using a pre-training stage (right) in concert with *STAMP*, which leads to lower loss compared with no pre-training stage (left).

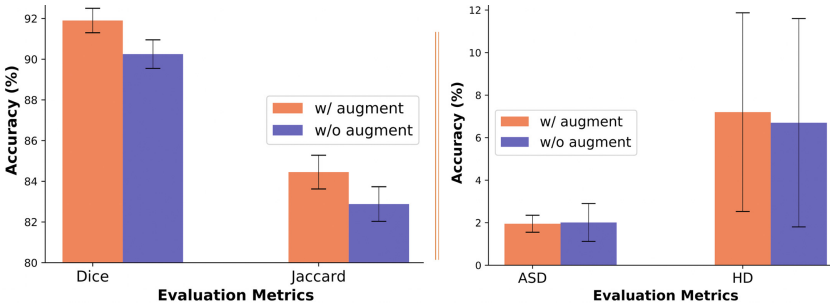

Fig. 8. Experiment conducted on a left atrial image datasets consisting of only 20% labeled data showing the benefits of using data augmentation (orange) in concert with STAMP, which leads to higher accuracy (Dice and Jaccard) compared with no data augmentation (purple). (Color figure online)

3.2.3 Effect of Data Augmentation

To improve generalization and significantly reduce error rate, we applied different strong and weak *data augmentation* strategies. Figure 8 shows a comparison of the model with and without the augmentation strategies. Our observation shows that when replacing weak augmentation with no augmentation, the model overfits the predicted unlabeled labels. The statistical significance of the *Dice and **Jaccard for STAMP model *with* and *without* data augmentation for 20% labeled data are represented by * and ** for $p-$values 0.1 and 0.001, respectively.

4 Conclusion

In this paper, we propose an effective Student-Teacher Augmentation-driven Meta pseudo-labeling (STAMP) model for 3D cardiac MRI image segmentation. The proposed framework mitigates the pseudo-labeling bias problem arising due to class imbalance by adopting a *threshold* where pixels with a confidence score higher than 0.5 will be used as pseudo labels, while the remaining are treated as ignored regions. Additionally, the proposed model also mitigates the overfitting challenge induced by the lack of a large pool of labeled data. The meta pseudo-labeling approach generates pseudo labels by a Teacher-Student mutual learning process where the Teacher learns from the Student's reward signal, which, in turn, best helps the Student's learning. Unlike the non-gradient exponential moving average (EMA) method, this reward signal is utilized to motivate the Teacher during the Student's learning process through the gradient descent algorithm. Moreover, the application of different strong and weak *data augmentation* strategies improve the generalization performance and reduce the error rate significantly. We evaluated our proposed framework within the SSL setting by comparing the segmentation results with those yielded by several existing methods. When using only 10% labeled data, STAMP achieves a 2.6 fold mean Dice improvement over the state-of-the-art RLSSS model. In addition, our proposed model outperforms existing methods in terms of both Jaccard and Dice, achieving 90.4% Dice and 82.7% Jaccard with only 10% labeled data and 91.9% Dice and 84.4% Jaccard with only 20% labeled data for atrial segmentation, both of which showed at least 2.6% improvement over the best methods and more than 11% improvement over fully-supervised traditional V-Net architecture.

References

1. Bai, W., et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 253–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_29
2. Balakrishnan, G., et al.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
3. Berthelot, D., et al.: Mixmatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems 32 (2019)

4. Chaitanya, K., et al.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Adv. Neural. Inf. Process. Syst. **33**, 12546–12558 (2020)
5. Chen, L.C., et al.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
6. DeVries, T., et al.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
7. French, G., et al.: Semi-supervised semantic segmentation needs strong, high-dimensional perturbations (2019)
8. Guo, H., et al.: Mixup as locally linear out-of-manifold regularization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3714–3722 (2019)
9. Iscen, A., et al.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5070–5079 (2019)
10. Laine, S., et al.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
11. Lee, D.H., et al.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896 (2013)
12. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_54
13. Liu, Y.C., et al.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
14. Milletari, F., et al.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
15. Oliver, A., et al.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems 31 (2018)
16. Ouali, Y., et al.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)
17. Pham, H., et al.: Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11557–11568 (2021)
18. Sajjadi, M., et al.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems 29 (2016)
19. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv. Neural. Inf. Process. Syst. **33**, 596–608 (2020)
20. Tarvainen, A., et al.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems 30 (2017)
21. Xie, Q., et al.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698 (2020)

22. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
23. Yun, S., et al.: CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
24. Zeng, X., et al.: Reciprocal learning for semi-supervised segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 352–361. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_33
25. Zheng, Q., et al.: Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. Med. Image Anal. **56**, 80–95 (2019)