*bioengineering*

MDPI

*Article*

# Cascaded Self-Supervision to Advance Cardiac MRI Segmentation in Low-Data Regimes

**Martin Urschler [1,2,*], Elisabeth Rechberger [3], Franz Thaler [1,3,4] and Darko Štern [2,4]**

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, 8036 Graz, Austria
[2] BioTechMed-Graz, 8010 Graz, Austria
[3] Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria
[4] Gottfried Schatz Research Center, Medical Physics and Biophysics, Medical University of Graz, 8036 Graz, Austria
* Correspondence: martin.urschler@medunigraz.at

## Abstract

Deep learning has shown remarkable success in medical image analysis over the last decade; however, many contributions focused on supervised methods which learn exclusively from labeled training samples. Acquiring expert-level annotations in large quantities is time-consuming and costly, even more so in medical image segmentation, where annotations are required on a pixel level and often in 3D. As a result, available labeled training data and consequently performance is often limited. Frequently, however, additional unlabeled data are available and can be readily integrated into model training, paving the way for semi- or self-supervised learning (SSL). In this work, we investigate popular SSL strategies in more detail, namely Transformation Consistency, Student–Teacher and Pseudo-Labeling, as well as exhaustive combinations thereof. We comprehensively evaluate these methods on two 2D and 3D cardiac Magnetic Resonance datasets (ACDC, MMWHS) for which several different multi-compartment segmentation labels are available. To assess performance in limited dataset scenarios, different setups with a decreasing amount of patients in the labeled dataset are investigated. We identify cascaded Self-Supervision as the best methodology, where we propose to employ Pseudo-Labeling and a self-supervised cascaded Student–Teacher model simultaneously. Our evaluation shows that in all scenarios, all investigated SSL methods outperform the respective low-data supervised baseline as well as state-of-the-art self-supervised approaches. This is most prominent in the very-low-labeled data regime, where for our proposed method we demonstrate 10.17% and 6.72% improvement in Dice Similarity Coefficient (DSC) for ACDC and MMWHS, respectively, compared with the low-data supervised approach, as well as 2.47% and 7.64% DSC improvement, respectively, when compared with related work. Moreover, in most experiments, our proposed method is able to greatly decrease the performance gap when compared to the fully supervised scenario, where all available labeled samples are used. We conclude that it is always beneficial to incorporate unlabeled data in cardiac MRI segmentation whenever it is present.

**Keywords:** semi-supervised learning; cardiac segmentation; self-training; student–teacher; pseudo-labeling



Academic Editors: Yuxiang Zhou and Xin Meng

Received: 11 July 2025
Revised: 6 August 2025
Accepted: 9 August 2025
Published: 12 August 2025

**Citation:** Urschler, M.; Rechberger, E.; Thaler, F.; Štern, D. Cascaded Self-Supervision to Advance Cardiac MRI Segmentation in Low-Data Regimes. *Bioengineering* **2025**, *12*, 872. https://doi.org/10.3390/bioengineering12080872

## 1. Introduction

Cardiovascular Diseases (CVDs) are the most common non-communicable diseases globally [1]. Their prevalence increased from 5.1% in 1990 to 6.8% in 2019, with a rising trend in almost all low-income countries as well as in some high-income countries. Further, CVDs are also the leading cause of death globally, most prominently due to ischemic heart disease, ischemic strokes, and intracerebral haemorrhages [1]. Assessment of CVD risk can be performed using biochemical, genetic, or imaging-based biomarkers. Examples for imaging biomarkers relevant for CVD include dysfunctionality of the left ventricle, plaque burden, and tissue composition [2], which can be measured and monitored non-invasively using ultrasound, computed tomography (CT) or magnetic resonance imaging (MRI). Especially MRI is often used to assess volume and function of left and right ventricle, either statically or dynamically during the heartbeat [3]. To derive such biomarkers from MRI, a multi-compartment segmentation is required, allowing to compute functional metrics like ejection fraction, wall thickness, ventricle volumes, and myocardial mass that aid in diagnosing disease and in planning treatments [4].

Due to entirely manual expert segmentation being tedious, costly, and time-consuming, fully automated supervised methods have been proposed for multi-compartment segmentation, predominantly using deep neural networks like the U-Net or nnU-Net model [5,6]. To perform supervised learning, images and corresponding ground truth expert segmentation labels have to be available. These datasets need to be sufficiently large to learn generalizable models; however, providing large labeled datasets for training is still tedious and costly. For instance, manual annotation of 3D whole heart segmentation labels as performed for the MMWHS dataset [7] require 6–10 h of interaction time per MR or CT volume [8]. Thus, in medical imaging practice, the amount of labeled data is often limited.

Whereas labeled data might be scarce, unlabeled data are often abundantly available, which motivates the use of semi- and self-supervised learning (SSL) methods [9,10]. Such approaches learn from a combination of labeled and unlabeled data and promise improved efficacy and a reduction in labeling effort, both generally for medical image analysis and also specifically for cardiac multi-compartment segmentation.

In this work, we explore which types of semi- and self-supervised techniques achieve the best performance for segmenting cardiac images while keeping the number of labeled images as low as possible. Therefore, we comprehensively compare three baseline strategies that are widely used in the literature, i.e., Transformation Consistency (TC), Student–Teacher (ST), and Pseudo-Labeling (PL) [11]. Moreover, we also exhaustively combine the three aforementioned SSL strategies, thus forming individual cascaded multi-step approaches to investigate. We study all baseline and combined SSL methods by comparing them with a supervised baseline as well as several recent state-of-the-art SSL methods from the literature on two cardiac datasets involving both 2D and 3D images. Furthermore, to assess the performance of SSL when the size of labeled training data is decreasing, we construct different setups with smaller and smaller numbers of subjects in the labeled training set. Finally, built upon the findings of our SSL combinations, we propose a novel multi-step PL algorithm based on cascaded ST stages that most effectively profits from labeled and unlabeled data. To the best of our knowledge, this is the first PL method which uses an ST cascade with strong input data augmentations. Our main contributions are summarized as follows:

1. We propose a novel SSL strategy for cardiac MRI segmentation by performing multi-step pseudo-labeling based on cascaded ST stages.
2. We extensively evaluate different individual SSL strategies and their combinations and compare them with our proposed cascaded algorithm on 2D and 3D imaging data.

3. We assess the performance of all investigated strategies in a low-data regime by systematically reducing the size of labeled data and comparing with the fully supervised case as well as related work.

## 2. Related Work

The taxonomy of Yang et al. [11] proposes a distinction of SSL approaches into several categories, involving hybrid, graph-based, generative, consistency regularization-based, and Pseudo-Labeling-based methods. In our work, we focus on consistency regularization and Pseudo-Labeling approaches as well as hybrid combinations between those, since they do not pose as much computational overhead as generative synthesis methods [12,13]. Moreover, the considered methods are inductive as opposed to graph-based methods which, according to [11], are mostly transductive and thus not well applicable in our cardiac segmentation scenario, where future test images are generally not available in advance.

### *2.1. Consistency Regularization*

Consistency regularization is based on the semi-supervised smoothness (cluster) assumption [9] and applies perturbations to the unlabeled data. Assuming the cluster assumption holds true, then perturbing data points within the same cluster should not change their predicted output semantically but solely perturb the output (segmentation labels) accordingly. The goal of consistency regularization is to minimize the discrepancy between the model output of an unlabeled data point and the accordingly perturbed model output of a transformed variant of the unlabeled data point. Transformations can be geometric (rotations, scaling, elastic deformation, etc.) as well as intensity-based (contrast variations, noise, blurring, etc.) [14]. Bortsova et al. proposed a Transformation Consistency approach for medical image analysis [15] using a Siamese network architecture with shared weights to segment chest X-ray images. Li et al. used transformations as well as dropout as a source of data perturbations to perform skin lesion segmentation of natural images [16]. Another line of work augments the input by masking parts of images and optionally combining them via mixing. Aiming to improve model robustness, DeVries et al. explored zero masking square parts of images for data augmentation in classification referred to as Cutout [17]. CutMix introduced by Yun et al. in [18] proposed to add random patches from the same images into the Cutout regions. French et al. then combined these two data augmentation techniques for consistency regularization in semantic segmentation [19].

An alternative manner to introduce consistency regularization is via multi-decoder consistency, where models have two or more branches which differ in their network architecture. Luo et al. implemented such an architecture simply by adding two additional distinct layers to a decoder, one for predicting segmentation outputs and the other to obtain a signed distance function representation of the output. They proposed their method as a dual task consistency (DTC) strategy, stating that the loss computes segmentations for labeled samples and additionally regresses signed distance functions for unlabeled samples [20]. Also, Wu et al. used two or three branches in their Mutual Consistency Network (MC-Net, MC-Net+) for left atrium segmentation with a common encoder and two or three separate decoders, where decoders differed in their upsampling strategies [21,22]. Three networks were also used by Huang et al., who proposed a main network and two auxiliary networks with varying skip connections in their decoders [23]. The outputs of the auxiliary models are passed through a sharpening function, thus serving immediately as a pseudo-label for the main model and the respective other auxiliary model.

In this study, we explore Transformation Consistency using strong data augmentation and make a comparison to various multi-decoder networks.

### *2.2. Student-Teacher*

When combining the predictions of several neural networks by forming an ensemble, aggregated predictions tend to be more accurate compared to single networks. This idea inspired Laine and Aila for their temporal ensembling work [24] by reusing networks at different (i.e., previous) training epochs to produce predictions for unlabeled data. This strategy was further developed by Tarvainen and Valpola to form the Mean Teacher or Student–Teacher model [25]. Different from [24], in a Student–Teacher model, the student and the teacher model are both used at training time, where predictions of the teacher model are used as training targets of unlabeled data from which the student learns via an unsupervised loss.

Yu et al. introduced an uncertainty aware Mean Teacher for 3D left atrium segmentation [26] via several forward passes involving dropout. This lead to the student only learning from targets where the teacher exceeded a certain confidence threshold. Wang et al. extended this approach by using uncertainty to interpolate between student and teacher predictions for every voxel [27]. A variant of a certainty-driven consistency loss for Student–Teacher was proposed in [28] based on filtering targets from the teacher using top-k certain predictions. Huang et al. implemented Student–Teacher consistency regularization for neuron segmentation in electron microscopy volumes [29]. A proxy task is used to pretrain the encoder weights of the student network to reconstruct the original sample from perturbed versions, which helps the Student–Teacher network in making most effective use of unlabeled samples. Lei et al. also used the Mean Teacher model for medical segmentation and added two discriminators for an additional adversarial loss [30]. The goal of the first discriminator is to evaluate the quality of segmentation results, whereas the second discriminator differentiates between the perturbed and the original unlabeled samples. By combining Mean Teacher with Mixup augmentation [31], Basak et al. encouraged networks to linearly interpolate between training samples, leading to a very simple yet effective strategy [32].

In this work, we study different variants of the Student–Teacher paradigm and compare them with hybrid methods.

### *2.3. Pseudo-Labeling*

There are two major directions in Pseudo-Labeling, self-training and deferring disagreement from different views or network models [11]. In self-training, an initially pretrained network (e.g., using reconstruction error) is used to produce predictions of unlabeled data which are then treated as the reference annotation for training another network on the pooled training samples [33,34]. Contrarily, Bai et al. train their initial network in a supervised manner on their cardiac MR segmentation task directly [35]. Introducing prediction confidence via ensembles and re-generating pseudo-labels regularly after a certain number of training iterations is used in [36] in the context of lung infection segmentation. To form pseudo-labels via disagreement of different models, Li et al. suggest to use several permutations of the input (three by three equally sized tiles) and take the average of the predictions for each modified input as pseudo-label [37]. Xie et al. propose Noisy Student, a Student–Teacher-like framework, where a teacher network is trained in a supervised manner and generates pseudo-labels for unlabeled samples [38]. Another very recent state-of-the-art line of work combines self-training with contrastive learning, which aims to bring feature representations of similar labeled images closer together while pushing feature representations of dissimilar labeled images further apart. Chaitanya et al. propose Local Contrastive Loss with Pseudo-Labels (LCLPL) for semantic medical image segmentation, where they employ a contrastive loss on the pixel-level through a separate decoding branch using the ground truth for labeled data and pseudo-labels for unlabeled data [39]. Close

to their work is the method proposed by Basak and Yin [40], who propose a patch-wise computation of contrastive loss instead of pixel-wise, which they call Pseudo-label Guided Contrastive Learning (PLGCL).

Pseudo-Labeling has been shown to be a promising SSL method recently [39,40]. Our evaluation demonstrates that combining the ideas of ST and PL can give a simple but very effective training scheme that is generally applicable for cardiac 2D and 3D MR segmentation and that achieves new state-of-the-art results.

## 3. Method

We design our study as a comprehensive evaluation of self-supervised strategies for cardiac MR image segmentation in low data regimes. Therefore, we investigate three methods: (i) Transformation Consistency, (ii) Student–Teacher without and with transformations, and (iii) Pseudo-Labeling. The following sections outline these methods and how we combine them, eventually leading to a novel cascaded self-supervised variant.

### *3.1. Transformation Consistency*

We build upon the Transformation Consistency work from [15], where every unlabeled sample $\mathbf{x}_u \in \mathbf{X}_u$ undergoes two different random transformations $TF_1(\mathbf{x}_u)$ and $TF_2(\mathbf{x}_u)$. The difference between the outputs for those two transformed samples forms the unsupervised loss component in their work. We modify their approach by solely using a single transformation $TF(\mathbf{x}_u)$ on unlabeled samples; see Figure 1 for an overview of the method. We use a set of potential geometric and intensity transformations inspired by [41], who apply the same transformations in the context of data augmentation for supervised multi-compartment cardiac segmentation. Each spatial transformation includes translation, rotation, scaling, and elastic deformation, with a uniformly sampled value within predefined ranges specifying the actual random transformation. For the intensity transformations, random shifting and scaling are chosen. Further details on the transformations can be found in Section 4.4.

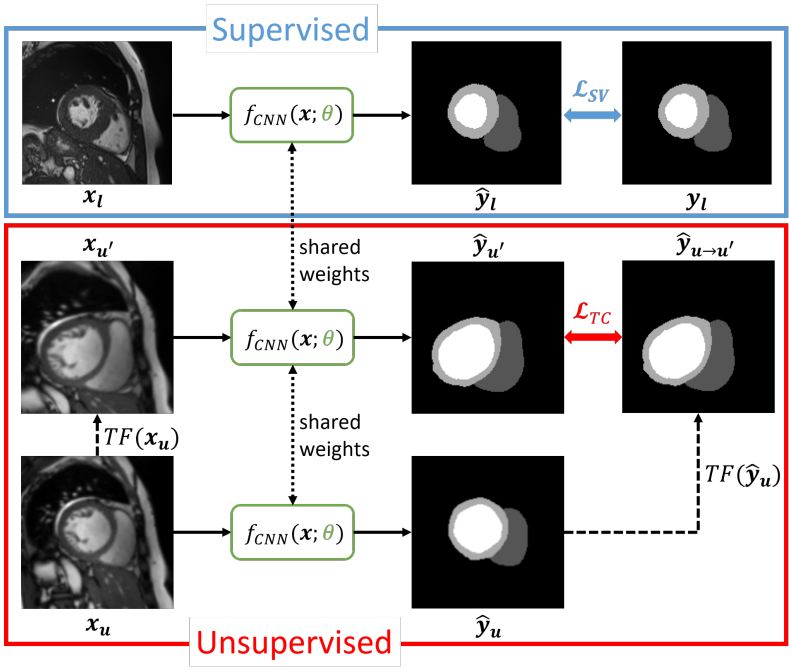



**Figure 1.** Transformation Consistency (TC) approach for self-supervised learning of cardiac multi-compartment segmentation. The supervised (SV) loss is accompanied by an unsupervised TC loss that penalizes deviations of differently transformed predictions. The segmentation CNN shares its weights between supervised and unsupervised components.

In our method, two samples are randomly drawn during each training iteration. One sample $\mathbf{x}_l$ is drawn from the set of labeled images $\mathbf{X}_l$ for computing the supervised loss component $\mathcal{L}_{SV}$, the other sample $\mathbf{x}_u$ is drawn from the set of unlabeled images $\mathbf{X}_u$ for computing the unsupervised loss component $\mathcal{L}_{TC}$. For the unsupervised path, the sample $\mathbf{x}_u$ is transformed to give $\mathbf{x}_{u'} = TF(\mathbf{x}_u)$. The two variants $\mathbf{x}_u$, $\mathbf{x}_{u'}$ are then forwarded through the single convolutional neural network (CNN) model $f(\mathbf{x};\theta)$, which produces corresponding predictions $\hat{\mathbf{y}}_u = f(\mathbf{x}_u;\theta)$ and $\hat{\mathbf{y}}_{u'} = f(\mathbf{x}_{u'};\theta)$. The CNN shares its weights between supervised and unsupervised components. To align the two predictions, $TF$ is applied to $\hat{\mathbf{y}}_u$, resulting in the target for the unsupervised loss $\hat{\mathbf{y}}_{u\to u'}$. Ideally, $\hat{\mathbf{y}}_{u'}$ and $\hat{\mathbf{y}}_{u\to u'}$ should be identical. Thus, the unsupervised TC loss is computed as the difference between $\hat{\mathbf{y}}_{u'}$ and $\hat{\mathbf{y}}_{u\to u'}$, in the form of the mean squared error (MSE):

$$\mathcal{L}_{TC} = MSE(\hat{\mathbf{y}}_{u'}, \hat{\mathbf{y}}_{u\to u'}). \tag{1}$$

To compute the supervised loss component $\mathcal{L}_{SV}$, we use a generalized Dice loss (GDL) [42], which compares predictions $\hat{\mathbf{y}}_l = f(\mathbf{x}_l;\theta)$ with their available ground truth segmentation $\mathbf{y}_l \in \mathbf{Y}_l$. Both losses, supervised and unsupervised, are then combined to form the total SSL loss via a weighted sum, weighting the Transformation Consistency loss component with a factor $\lambda$:

$$\mathcal{L}_{SSL,TC} = \mathcal{L}_{SV} + \lambda\mathcal{L}_{TC}. \tag{2}$$

After the losses for the training are computed, the back-propagation of the gradient takes place. As $\hat{\mathbf{y}}_{u\to u'} = TF(f(\mathbf{x}_u;\theta))$ serves as the unsupervised target for $\hat{\mathbf{y}}_{u'}$, $\hat{\mathbf{y}}_{u\to u'}$ is excluded from back-propagation by treating it like a constant to avoid gradient collapse as suggested in [43].

### *3.2. Student–Teacher*

Motivated by the work of [25], we implement two ST models, where the teacher provides targets for the student model and in turn shares its knowledge via an exponential moving average (EMA).

In each ST training iteration, a labeled sample $\mathbf{x}_l \in \mathbf{X}_l$ and an unlabeled sample $\mathbf{x}_u \in \mathbf{X}_u$ are randomly drawn. The computation of the supervised loss is based on forwarding each labeled sample $\mathbf{x}_l \in \mathbf{X}_l$ through the student model only. Its output, $\hat{\mathbf{y}}_l = f(\mathbf{x}_l;\theta_S))$ is compared with the ground truth segmentation $\mathbf{y}_l$ via the GDL, same as for the Transformation Consistency method. To compute the unsupervised ST loss component $\mathcal{L}_{ST}$, the teacher prediction $\hat{\mathbf{y}}_T$ serves as a target for the prediction of the student $\hat{\mathbf{y}}_S$. Both losses are combined to form the total SSL loss:

$$\mathcal{L}_{SSL,ST} = \mathcal{L}_{SV} + \lambda\mathcal{L}_{ST}. \tag{3}$$

After the back-propagation of the total SSL loss through the student network is performed, the weights of the student $\theta_S$ are updated. Conversely, the weights of the teacher network $\theta_T$ are not modified by back-propagation directly; instead, they are defined as the EMA of the student weights over all gradient updates and can be recursively defined as

$$\theta_T^i = \alpha\theta_T^{i-1} + (1-\alpha)\theta_S^i. \tag{4}$$

We evaluate the ST loss component in two variants, leading to two distinct SSL methods.

#### 3.2.1. Student–Teacher Without Transformations

Here, the same input $\mathbf{x}_u$ serves as input to the student model $f(\mathbf{x}_u;\theta_S)$ and the teacher model $f(\mathbf{x}_u;\theta_T)$. Both models use dropout during training for regularization. The MSE

between the student predictions $\hat{\mathbf{y}}_S = f(\mathbf{x}_u; \theta_S)$ and the target as predicted from the teacher $\hat{\mathbf{y}}_T = f(\mathbf{x}_u; \theta_T)$ gives the unsupervised loss component $\mathcal{L}_{ST}$:

$$\mathcal{L}_{ST} = MSE(\hat{\mathbf{y}}_S, \hat{\mathbf{y}}_T). \tag{5}$$

#### 3.2.2. Student–Teacher with Transformations

For the second ST variant (see Figure 2 for an illustration), Transformation Consistency as explained in Section 3.1 is used to additionally augment the unlabeled inputs. A random transformation $TF$ is applied in each training iteration on a sample $\mathbf{x}_u$. While the teacher $f(\mathbf{x}; \theta_T)$ receives the unmodified $\mathbf{x}_u$ as input, $TF(\mathbf{x}_u)$ is forwarded to the student model $f(\mathbf{x}; \theta_S)$. The student outputs the prediction $\hat{\mathbf{y}}_S$, whereas the teacher predicts $\hat{\mathbf{y}}_T$. Same as for tTransformation Consistency, the predictions are aligned by applying $TF$ on $\hat{\mathbf{y}}_T$ to obtain $\hat{\mathbf{y}}_{T\to S} = TF(\hat{\mathbf{y}}_T)$. The unsupervised loss $\mathcal{L}_{ST}$ is then computed as MSE between $\hat{\mathbf{y}}_{T\to S}$ and $\hat{\mathbf{y}}_S$:

$$\mathcal{L}_{ST} = MSE(\hat{\mathbf{y}}_S, \hat{\mathbf{y}}_{T\to S}). \tag{6}$$

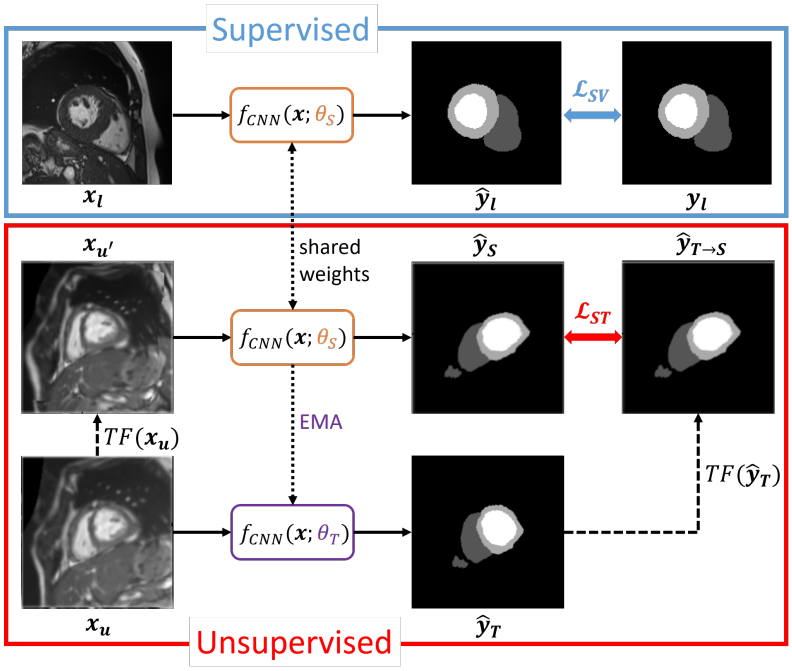



**Figure 2.** Student–Teacher approach with the use of Transformation Consistency for self-supervised learning. Different from pure TC, the teacher network weights are computed as an exponential moving average (EMA) of the student weights. Differently transformed predictions from student and teacher are penalized using the unsupervised ST loss component.

### *3.3. Self-Training via Pseudo-Labeling*

For Pseudo-Labeling variants, the training process consists of three consecutive stages. Firstly, the model $f(\mathbf{x}; \theta_1)$ is trained either based on the subset of supervised samples alone (see Figure 3) or by using one of the two previously discussed SSL strategies (TC or ST). Secondly, for every unlabeled sample $\mathbf{x}_u \in \mathbf{X}_u$, pseudo-labels are generated in an inference stage, leading to the set $\mathbf{Y}_{PL}$. Finally, another model $f(\mathbf{x}; \theta_2)$ is trained from scratch with randomly initialized weights, but now in a purely supervised (SV) manner, with the union of $\mathbf{Y}_l$ and $\mathbf{Y}_{PL}$ as the set of target labels. For each iteration of the third stage, we randomly draw two samples, one sample $\mathbf{x}_l$ from the labeled set $\mathbf{X}_l$, the other sample $\mathbf{x}_u$ from the originally unlabeled set $\mathbf{X}_u$, both with their respective target label. This is performed to ensure that samples from the originally labeled set have sufficient influence during model training, as the number of pseudo-labeled samples is potentially much larger and

their pseudo-label-based segmentations are expected to have a lower quality compared to samples for which actual ground truth segmentations are available. We note that we do not perform any filtering or selection of pseudo-labels but use all available pseudo-labels for $\mathbf{Y}_{PL}$. The supervised loss in the final stage is composed of two different supervised losses, which are both implemented with a GDL. We separate the two losses to introduce a weighting factor for controlling the influence of the samples drawn with pseudo-label targets, leading to the total pseudo-labeling loss:

$$\mathcal{L}_{SSL,PL} = \mathcal{L}_{SV} + \lambda \mathcal{L}_{PL}, \tag{7}$$

where $\mathcal{L}_{SV}$ penalizes discrepancies between predictions $\hat{\mathbf{y}}_l = f(\mathbf{x}_l; \theta_2)$ and targets $\mathbf{y}_l \in \mathbf{Y}_l$ and $\mathcal{L}_{PL}$ penalizes discrepancies between $\hat{\mathbf{y}}_u = f(\mathbf{x}_u; \theta_2)$ and targets $\mathbf{y}_{PL} \in \mathbf{Y}_{PL}$.

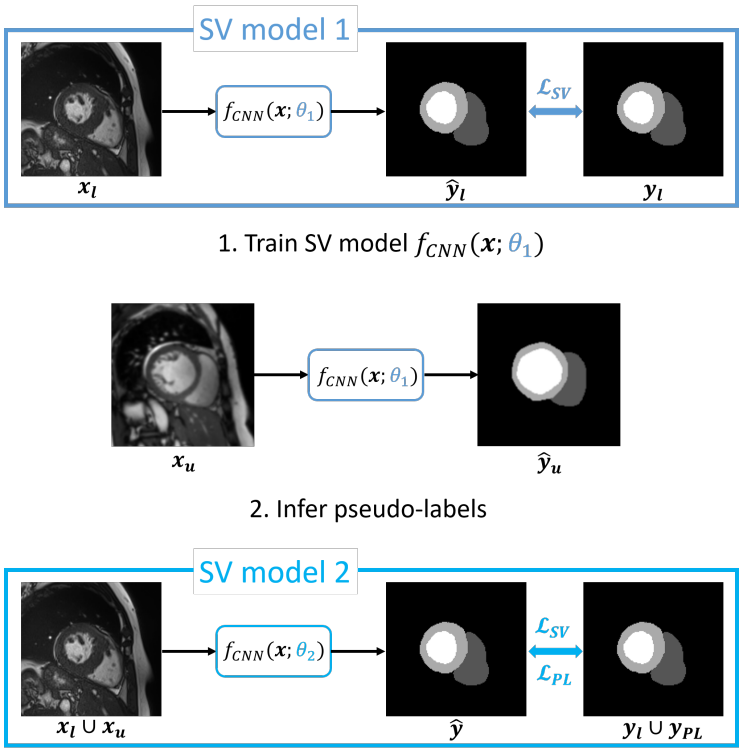



**Figure 3.** Traditional Pseudo-Labeling method based on two supervised training rounds. Model 1 trained in the first round is used in the second step to infer pseudo-labels for all unlabeled samples. In the second training round, the final prediction Model 2 is trained via a combination of supervised losses, which are computed using expert annotated ground truth segmentations for labeled data and the generated pseudo-labels for unlabeled data.

### *3.4. Cascaded Self-Supervision*

Our main methodological contribution is to combine the idea of Student–Teacher with Pseudo-Labeling, thus creating a cascade of self-trained SSL approaches (see Figure 4). We hypothesize that such a cascade achieves a best of both worlds approach for segmentation, combining supervised and unsupervised components with as much training data as possible. Firstly, we train a model with Student–Teacher based on an unsupervised and a supervised set of samples, as described in Section 3.2. Secondly, we use the trained SSL model to infer pseudo-labels $\mathbf{Y}_{PL}$ for all $\mathbf{x}_u \in \mathbf{X}_u$. We then combine the predictions $\mathbf{Y}_{PL}$ with the ground truth labels $\mathbf{Y}_l$, thus forming a label set $\mathbf{Y}_{full} = \mathbf{Y}_l \cup \mathbf{Y}_{PL}$. Different from before, in the third stage, we train a new Student–Teacher model, where labeled samples $\mathbf{x}_l$ are

now drawn from $\mathbf{X}_{full} = \mathbf{X}_l \cup \mathbf{X}_u$ due to the availability of pseudo-labels in $\mathbf{Y}_{full}$. Samples $\mathbf{x}_u$ are also drawn from $\mathbf{X}_{full}$ but ignoring any ground truth labels. To balance labeled and unlabeled images such that pseudo-labeled images do not dominate during training, we draw one sample of each category per training iteration. Thus, we achieve a cascade of two semi-supervised Student–Teacher models, which makes optimal use of labeled, pseudo-labeled, and unlabeled samples simultaneously. The loss function in the third stage uses a weighted combination of GDL terms for the labeled ($\mathcal{L}_{SV}$) and pseudo-labeled samples ($\mathcal{L}_{PL}$), as well as a weighted MSE term for the unlabeled samples ($\mathcal{L}_{ST}$):

$$\mathcal{L}_{SSL,cascade} = \mathcal{L}_{SV} + \lambda_1 \mathcal{L}_{PL} + \lambda_2 \mathcal{L}_{ST}. \tag{8}$$

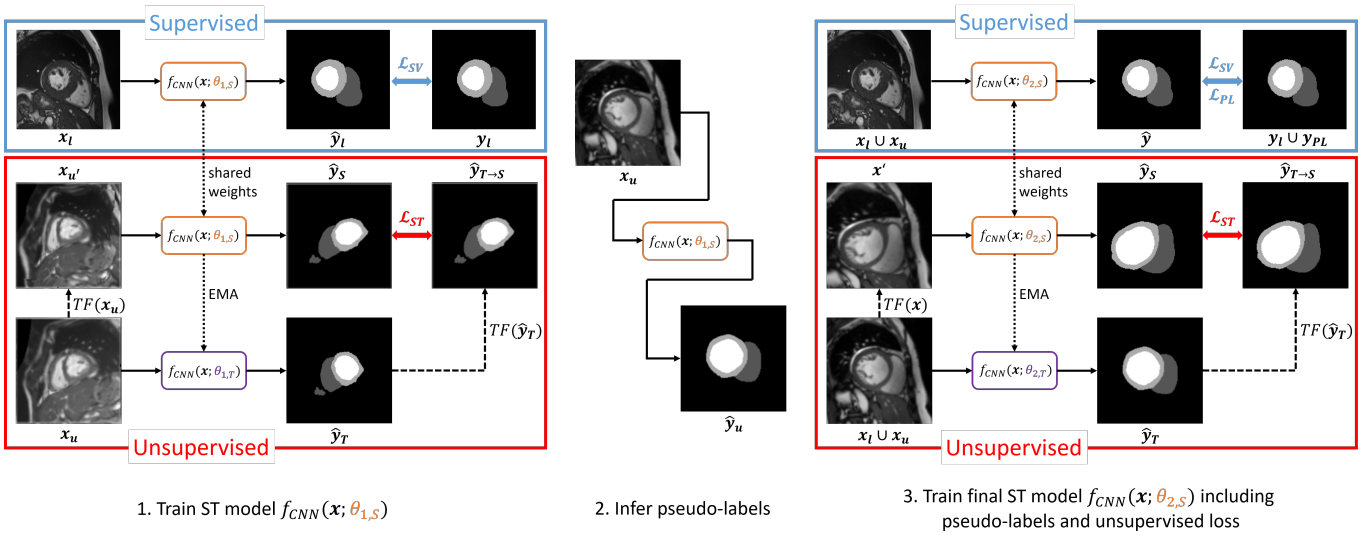
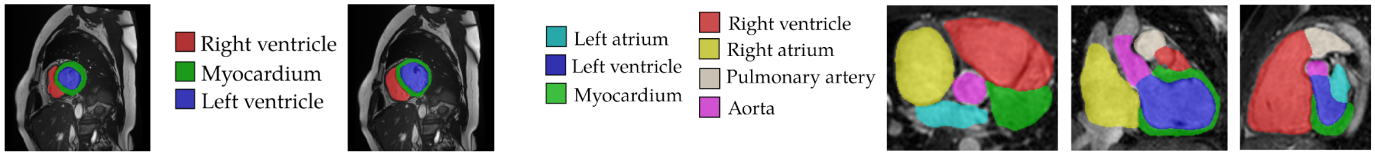



**Figure 4.** Proposed cascaded Self-Supervision method which employs Pseudo-Labeling and introduces a self-supervised cascaded Student–Teacher model, i.e., using self-training in the first and second training rounds ($ST_{TC}$-PL-$ST_{TC}$). Both ST models benefit from labeled and unlabeled samples during training. In addition, the second ST model is trained using pseudo-labels for the unsupervised set obtained through the first ST model. This results in two supervised and one unsupervised loss components.

### *3.5. Datasets*

We use two cardiac segmentation datasets for evaluating our semi-supervised self-training methods. The ACDC dataset consists of 4D Cine MR images, which capture morphological changes of the heart during the heartbeat [44]. Three labels are available for this dataset, namely myocardium (MYO), right ventricle (RV), and left ventricle (LV); see Figure 5a. Data acquisition was synchronized with an ECG, using an SSFP sequence in short axis orientation. While the exact numbers vary, on average, each 4D sample in the dataset consists of about 26 time steps and thus around 26 3D volumes that represent the cardiac cycle. Ground truth segmentations by medical experts are only acquired for the systolic and diastolic phase of the cardiac cycle, while the 3D volumes of the remaining time steps remain unlabeled. The 3D volumes themselves capture roughly 10 2D short-axis slices on average that cover the heart from base to apex with a significant slice thickness of 5 to 10 mm, thus leading to severe anisotropy. In comparison, the spatial in-plane resolution ranges from 1.37 to 1.68 $mm^2$. Furthermore, the 2D short-axis slices of the ACDC dataset are in some cases misaligned due to motion during data acquisition. Due to the large slice thickness as well as the misaligned slices, we use the ACDC dataset in 2D, which is the same setup as in related work. The whole dataset comprises image data from 150 patients, divided into five evenly sized subgroups: one healthy group and four with cardiac diseases

(myocardial infarction, dilated and hypertrophic cardiomyopathy, and abnormal ventricle). Of these 150 patients, 100 are part of the training set, while the remaining 50 patients belong to the official testing dataset for which we did not have access to the ground truth annotations. Consequently, for training and testing of the evaluated methods, only the training set with 100 patients is used in cross-validation.




(**a**) (**b**)

**Figure 5.** Datasets used for evaluating our self-training approaches in the cardiac MRI setting. (**a**) ACDC dataset [44] consisting of dynamically acquired 2D slices of the heart and providing a three-label annotation. (**b**) MMWHS dataset [7] consisting of 3D MR volumes with high spatial resolution and providing a seven-label annotation.

## 4. Experimental Setup

As a second dataset, we use the MR data from the MMWHS challenge held in conjunction with MICCAI 2017 [7]. Specifically, the dataset consists of 60 MR heart volumes in 3D, covering the region from the upper abdomen to the aortic arch. While the training set consists of 20 samples with ground truth labels, the test set encompasses 40 samples for which ground truth segmentations are not publicly available. The free-breathing MR volumes were acquired with a navigator-gated balanced SSFP sequence, with a nearly isotropic resolution of approximately $(0.8-1) \times (0.8-1) \times (1-1.6)$ mm after the reconstruction. The segmentation labels include seven cardiac substructures, also shown in Figure 5b. These are the blood cavities, i.e., left ventricle (LV), right ventricle (RV), left atrium (LA) and right atrium (RA), the myocardium (MYO) of the left ventricle, the aorta (AO) starting from the aortic valve to the upper part of the atria, and the pulmonary artery (PA) including the pulmonary valve up to the bifurcation point.

### *4.1. Data Preprocessing*

ACDC: Due to the anisotropic resolution of the ACDC dataset, we extract the 2D slices from each time step. With about 26 time steps and ten slices in the out-of-plane dimension on average, this results in around 260 2D images per patient. Two patients (IDs: 94, 88) were removed from the dataset, as their out-of-plane dimension was defined as superior to inferior, in contrast to the out-of-plane dimension for all other cases, which was defined as left to right. Additionally, slices were excluded from the dataset in case one or several of the labels covered less than 0.07% of the image in one of the two labeled time steps, since these slices introduced ambiguities due to inconsistencies in ground truth labeling of heart base and apex. In total, 25297 image slices of the ACDC dataset remained, of which approximately 7% (1670 slices) are labeled.

MMWHS and ACDC: To guarantee a common intensity range for the MR image intensities, robust normalization of MR images was performed by mapping the 95th percentile of intensity values per image to $-1$ and 1, respectively. Further, the intensity values for images and labels were preprocessed with Gaussian smoothing using a standard deviation of $\sigma = 1$. Images and label masks are resampled to a size of $160 \times 160$ pixel with a spacing of $1 \times 1$ mm for ACDC slices and $96 \times 96 \times 96$ voxel with $2 \times 2 \times 2$ mm for MMWHS volumes to ensure a constant input size and resolution for the network. For resampling, we use linear interpolation for images and nearest neighbor interpolation for label masks.

### *4.2. Data Augmentation*

To increase the diversity of the available images, we employ strong on-the-fly data augmentation during training using the framework of Payer et al. [41]. Both labeled and unlabeled samples are augmented using random spatial transformations. Sampled from a uniform distribution independently per dimensions, these include translation between $[-20, 20]$ pixels, rotation ranging from $[-0.35, 0.35]$ radians, and scaling with a factor ranging from $[-20, 20]$ pixels. Additionally, the unlabeled samples are transformed with randomized elastic deformations, with a maximum deformation value of 15 pixels and eight grid nodes for each dimension. Intensity transformations include random global shifting sampled from $[-0.2, 0.2]$ and global scaling ranging from $[-0.4, 0.4]$. Due to the characteristic intensity distribution of MR images, we set intensities below $-1$ to $-1$.

### *4.3. Neural Network Architecture*

We use a U-Net like architecture [5] to implement all our segmentation CNNs $f(\mathbf{x}; \theta)$. The network consists of a contracting and an expanding path, as well as skip connections between them. The contracting and the expanding path consist of four blocks each. In both paths, one such block is composed of two convolution layers with zero-padding and 64 filter channels, as well as an intermediate dropout layer [45] with a dropout rate of 0.1 [46,47]. For the 3D inputs of the MMWHS dataset, we use a $3 \times 3 \times 3$ kernel for convolution layers and a $3 \times 3$ kernel in the case of the 2D inputs of the ACDC dataset. Each convolution layer is followed by leaky ReLU as activation function with a slope of 0.1. After each block in the contracting path, we use average pooling for downsampling. Complementary, after each block in the expanding path, we employ linear upsampling, both with a factor of two. The skip connections concatenate intermediate features from the end of a block in the contracting path to the feature dimension of the matching level in the expanding path at the beginning of a block. Lastly, we employ softmax as the final activation function of the whole architecture to receive a pixel-wise probability distribution for each segmentation class before computing the losses.

### *4.4. Implementation Details*

The weights of the convolution layers are initialized according to the method proposed in [48]. For each iteration during supervised training, we randomly draw one sample $\mathbf{x}_l$ from the set of labeled samples $\mathbf{X}_l$. When training semi-supervised methods, we additionally draw one sample $\mathbf{x}_u$ from the set of unlabeled samples $\mathbf{X}_u$ per iteration to balance the influence of labeled and unlabeled samples. The learning rate follows an exponential decay with a rate of 0.1. In the case of SV and TC training, the initial learning rate is $1 \times 10^{-4}$, for ST and PL experiments the initial learning rate is set to $5 \times 10^{-5}$. As optimizer, Adam [49] is selected with the decay rate parameters defined as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For ST, the EMA parameter $\alpha$ for the weight update of the teacher model is chosen as 0.999. For the consistency weighting factor $\lambda$ for TC and ST, the factor 10 is chosen, whereas for PL training, the PL weighting factor is set to 1. For the cascaded ST method, $\lambda_1$ is set to 1 and $\lambda_2$ is set to 10. All weighting parameter choices are performed empirically, in a phase of initial experiments. The neural networks are trained for a total number of 180,000 iterations on the ACDC data and 40,000 iterations in the case of MMWHS. In all our self-training experiments, we train models for the first half of iterations in a supervised manner only. Only after this pre-training stage, the unsupervised learning paths are added to the training scheme, thus adding the influence of the unlabeled data. All parameters are selected after an initial empirical trial stage and according to prior experience in our group [41,46,50].

### *4.5. Evaluation Metrics*

We evaluate the performance of all considered methods by computing the Dice Similarity Coefficient (DSC) in percent as well as the Average Symmetric Surface Distance (ASSD) in mm. All metrics are computed after linearly resampling the prediction $\hat{\mathbf{y}}$ to its respective original spacing and dimension to allow a fair comparison. For multi-label segmentation, the DSC metric is defined as the average of the overlaps for all labels $c \in C$, i.e.,

$$DSC(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|C|} \sum_{c \in C} \frac{2|\mathbf{y}_c \cap \hat{\mathbf{y}}_c|}{|\mathbf{y}_c + \hat{\mathbf{y}}_c|}, \tag{9}$$

where $\mathbf{y}$ refers to the ground truth segmentation and $\hat{\mathbf{y}}$ to the predicted segmentation. Complementary to the overlap-oriented DSC metric, the ASSD assesses the segmentation performance from a boundary-oriented perspective. The ASSD metric is defined as

$$ASSD(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{|C|} \sum_{c \in C} \frac{\sum_{y_c \in \mathbf{y}_c} d(y_c, \hat{\mathbf{y}}_c) + \sum_{\hat{y}_c \in \hat{\mathbf{y}}_c} d(\hat{y}_c, \mathbf{y}_c)}{|\mathbf{y}_c| + |\hat{\mathbf{y}}_c|}, \tag{10}$$

where $d(\cdot)$ is the Euclidean distance of a point $a$ to the closest point $b$ in a set of points $\mathbf{b}$, i.e.,

$$d(a, \mathbf{b}) = \min_{b \in \mathbf{b}} ||a - b||. \tag{11}$$

For all our internal comparison experiments, scores are presented with their mean and standard deviation $\mu \pm \sigma$ and each experiment was repeated three times for every cross-validation fold. The standard deviation was first averaged over all fold repetitions for each sample and the final reported $\sigma$ was computed over all cross-validation folds.

### *4.6. Self-Training Method Variants*

In our comprehensive evaluation of self-training methods, we compare 10 different variants for both datasets, ACDC and MMWHS, with an increasing number of labeled samples in the supervised set, thus totaling four data regimes. There are four baseline methods, i.e., supervised (SV), Transformation Consistency (TC), Student–Teacher without TC ($ST_{noTC}$), and Student–Teacher with TC ($ST_{TC}$). Then, we derive three traditional supervised PL variants, where initial prediction models were trained either in a supervised manner (SV-PL-SV) or in a self-training manner (TC-PL-SV, $ST_{TC}$-PL-SV). Finally, we explore our proposed cascaded self-training combination $ST_{TC}$-PL-$ST_{TC}$ and also investigate two ablation versions, i.e., $ST_{noTC}$-PL-$ST_{noTC}$, as well as SV-PL-$ST_{TC}$.

### *4.7. Training Setup*

Internal evaluation: To evaluate the effectiveness of our studied self-supervised methods, we created different data setups with an increasing size of patients with available ground truth annotations, but a constant size of unlabeled patients.

For ACDC, in each cross-validation fold, the image slices of 75 patients from the original training set were chosen as the per fold training dataset. The unlabeled set $\mathbf{X}_u$ for ACDC is always composed of the image slices of all 75 patients from that fold. Contrarily, the labeled set $\mathbf{X}_l$ consists either of 5 (7%), 15 (20%), 25 (33%) or all 75 (100%) labeled patients, giving our four supervised setups. Note that ground truth segmentations are only available for 2 out of the 25 time steps per patient, and consequently only the 2D slices of these two time steps are considered to be part of the labeled dataset (see Section 3.5). We use a four-fold cross-validation for each of the four setups; thus, there are always the labeled slices from 25 patients in the held out test set. To ensure that slices from one patient stay within training or test set of a fold, patient IDs were shuffled randomly across the splits as opposed to shuffling image slices.

For our internal evaluation of the MMWHS dataset, again, four different supervised setups with varying amounts of labeled patient samples were defined in a three-fold cross-validation, i.e., setups with 3 (20%), 5 (35%), 7 (50%), and 14 (100%) patient volumes out of the 20 potentially available annotated samples (6 were kept for respective test sets in the cross-validation). Thus, all setups contained 54 samples in total, which were used as unsupervised set $\mathbf{X}_u$. Same as for ACDC, labeled samples were contained in the unsupervised set: $\mathbf{X}_l \subseteq \mathbf{X}_u$.

Comparison to recent related work: To also fairly compare the results of our proposed methods to the LCLPL work from [39], their evaluation setup was exactly reproduced for both MMWHS and ACDC. In contrast to our work, which employs the MMWHS data in 3D with size a $96 \times 96 \times 96$ pixels and a physical resolution of $2 \times 2 \times 2$ mm [39], image slices were extracted from the MMWHS volumes with a size of $160 \times 160$ pixels and a physical resolution of $1.5 \times 1.5$ mm. The fixed test set includes 20 patients for ACDC and 10 for MMWHS. For both ACDC and MMWHS datasets, three different supervised setups were constructed, containing the data of solely 1, 2, or 8 labeled patients each. This lead to percentages of 2%, 4%, or 16% for ACDC, as well as 10%, 20%, or 80% for MMWHS, respectively. The unlabeled sets comprised 10 patients for MMWHS and 52 for ACDC, respectively. The labeled patient images were also part of the unlabeled set: $\mathbf{X}_l \subseteq \mathbf{X}_u$. An entirely supervised baseline benchmark setup was also given, which comprised 78 ACDC patients or 10 MMWHS patients, i.e., (100%), respectively.

The second comparison was performed with PLGCL from Basak et al. [40], following the same setup from [51] for ACDC. In their test set, 20 patients were included. Another 70 different patients were used in each of the two different supervised setups, which included either 7 (10%) or 14 (20%) labeled patients. In their setup, these labeled patients were not part of the unsupervised set: $\mathbf{X}_l \not\subseteq \mathbf{X}_u$. Also, an entirely supervised baseline benchmark setup was used, consisting of all 70 (100%) labeled patients. The unlabeled set consisted of 70 different patients from the dataset.

## 5. Results and Discussion

We provide the results on the ACDC dataset in Table 1 and results for the MMWHS are given in Table 2. Our most promising self-training SSL method $ST_{TC}$-PL-$ST_{TC}$ as well as the corresponding baseline method $ST_{TC}$ are also compared to recent related methods. Our quantitative results when comparing to LCLPL from Chaitanya et al. [39] can be found in Table 3, whereas results when comparing to PLGCL [40] can be found in Table 4. Lastly, we also show some qualitative results for ACDC in Figure 6 as well as for MMWHS in Figure 7.

**Table 1.** Results from our internal evaluation comparing nine SSL methods and the supervised (SV) baseline on 2D ACDC dataset. Four different cross-validation setups were used for each percentage of patients in the labeled set, ranging from the lowest data regime of 7% up to 100%. Presented are the mean and standard deviation of DSC in % and ASSD in mm over three repetitions of the experiments. Best scores for mean and standard deviation are bold, second best scores are underlined.

| Method | ACDC: Percentage of Patients in Labeled Set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 7% | | 20% | | 33% | | 100% | |
| | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ |
| SV | 78.26 ± 19.22 | 2.15 ± 2.35 | 86.57 ± 14.01 | 1.19 ± 1.40 | 87.14 ± 11.22 | 0.99 ± 1.05 | 89.65 ± 7.80 | 0.92 ± 1.06 |
| TC | 84.18 ± 12.67 | 1.54 ± 1.68 | 88.01 ± 10.23 | 1.11 ± 1.33 | 88.81 ± 9.34 | 1.02 ± 1.24 | 89.73 ± 7.43 | 0.92 ± 1.03 |
| $ST_{noTC}$ | 85.19 ± 12.54 | 1.26 ± 1.39 | 88.98 ± 9.64 | 0.91 ± 1.14 | 89.69 ± 8.69 | 0.84 ± 0.96 | 90.62 ± 7.24 | 0.75 ± 0.86 |
| $ST_{TC}$ | 86.90 ± 10.86 | 1.14 ± 1.20 | 89.48 ± 8.88 | 0.87 ± 0.99 | 89.88 ± 8.54 | **0.83 ± 0.88** | **90.81** ± 6.88 | **0.75 ± 0.83** |
| SV-PL-SV | 85.72 ± 11.24 | 1.39± 1.67 | 88.98 ± 8.65 | 1.00 ± 1.19 | 89.44 ± 8.34 | 0.92 ± 1.03 | 89.93 ± 7.29 | 0.86 ± 0.91 |
| TC-PL-SV | 86.55 ± 10.20 | 1.27 ± 1.41 | 88.89 ± 9.08 | 0.99 ± 1.09 | 89.43 ± 8.39 | 0.94 ± 1.16 | 89.90 ± 7.43 | 0.87 ± 0.94 |
| $ST_{TC}$-PL-SV | 87.57 ± 9.77 | 1.15 ± 1.27 | 89.31 ± 8.67 | 0.95 ± 1.12 | 89.66 ± 8.07 | 0.90 ± 1.05 | 90.12 ± 7.30 | 0.86 ± 1.00 |

**Table 1.** *Cont.*

| Method | ACDC: Percentage of Patients in Labeled Set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 7% | | 20% | | 33% | | 100% | |
| | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ |
| SV-PL-$ST_{TC}$ | 86.56 ± 10.62 | 1.17 ± 1.36 | 89.37 ± 8.22 | 0.91 ± 1.02 | 89.83 ± **7.38** | 0.86 ± 0.93 | 90.18 ± 6.99 | 0.81 ± 0.86 |
| $ST_{noTC}$-PL-$ST_{noTC}$ | 87.10 ± 9.51 | 1.13 ± 1.18 | 89.02 ± 8.74 | 0.95 ± 1.08 | 89.53 ± 7.76 | 0.90 ± 0.99 | 89.89 ± 7.32 | 0.84 ± 0.94 |
| $ST_{TC}$-PL-$ST_{TC}$ | **88.43** ± **8.90** | **1.02** ± **1.08** | **89.82** ± **7.96** | **0.86** ± **0.96** | **90.04** ± 7.43 | 0.85 ± 0.93 | 90.49 ± **6.82** | 0.79 ± 0.85 |

**Table 2.** Results from our internal evaluation comparing nine SSL methods and the supervised (SV) baseline on 3D MMWHS dataset. Three different cross-validation setups were used for each percentage of patients in the labeled set, ranging from the lowest data regime of 20% up to 100%. Presented are the mean and standard deviation of DSC in % and ASSD in mm over three repetitions of the experiments. Best scores for mean and standard deviation are bold, second best scores are underlined.

| Method | MMWHS: Percentage of Patients in Labeled Set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 20% | | 35% | | 50% | | 100% | |
| | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ | DSC (%) ↑ | ASSD (mm) ↓ |
| SV | 81.44 ± 6.02 | 3.12 ± 1.78 | 85.04 ± 5.01 | 2.45 ± 1.84 | 85.97 ± 5.13 | 2.06 ± 1.38 | 87.71 ± 3.69 | 1.57 ± 0.90 |
| TC | 85.72 ± 3.20 | 1.65 ± 0.53 | 87.29 ± 2.97 | 1.44 ± 0.50 | 87.63 ± 3.35 | 1.42 ± 0.55 | 88.36 ± 3.07 | 1.30 ± 0.46 |
| $ST_{noTC}$ | 84.08 ± 4.03 | 2.00 ± 0.79 | 87.17 ± 3.24 | 1.51 ± 0.59 | 88.01 ± 3.27 | 1.40 ± 0.58 | 88.84 ± 2.83 | 1.22 ± 0.42 |
| $ST_{TC}$ | 86.14 ± 2.92 | 1.55 ± 0.42 | 87.56 ± 2.86 | 1.40 ± 0.48 | 88.21 ± 2.96 | 1.32 ± 0.46 | 88.69 ± 2.91 | 1.24 ± 0.43 |
| SV-PL-SV | 85.19 ± 4.75 | 2.08 ± 1.04 | 87.74 ± 3.20 | 1.41 ± 0.57 | 88.71 ± 3.45 | 1.27 ± 0.52 | 89.30 ± 2.82 | 1.18 ± 0.43 |
| TC-PL-SV | 87.58 ± **2.61** | 1.36 ± 0.40 | 88.50 ± 2.72 | 1.27 ± 0.45 | 89.06 ± 2.87 | 1.20 ± 0.42 | 89.25 ± 3.04 | 1.19 ± 0.46 |
| $ST_{TC}$-PL-SV | 87.68 ± 2.72 | 1.36 ± 0.43 | 88.55 ± 2.66 | 1.29 ± 0.47 | 89.17 ± 2.68 | 1.22 ± 0.42 | 89.42 ± 2.57 | 1.15 ± 0.38 |
| SV-PL-$ST_{TC}$ | 85.60 ± 4.73 | 2.02 ± 1.09 | 88.18 ± 3.17 | 1.32 ± 0.53 | 89.30 ± 2.94 | 1.15 ± 0.44 | **90.04** ± 2.48 | **1.05** ± 0.31 |
| $ST_{noTC}$-PL-$ST_{noTC}$ | 85.67 ± 4.25 | 1.76 ± 0.80 | 87.54 ± 3.21 | 1.35 ± 0.49 | 89.12 ± 2.73 | 1.17 ± 0.41 | 89.76 ± 2.49 | 1.10 ± 0.35 |
| $ST_{TC}$-PL-$ST_{TC}$ | **88.16** ± 3.02 | **1.27** ± **0.41** | **89.09** ± **2.53** | **1.18** ± **0.40** | **89.72** ± **2.61** | **1.07** ± **0.36** | 90.00 ± **2.42** | 1.06 ± **0.31** |

**Table 3.** Results from comparison of our proposed method with Chaitanya et al. [39] on 2D ACDC and 3D MMWHS datasets, using their exact evaluation setup. Four percentages of patients in the labeled set were investigated. Mean DSC performance measures of related methods were taken from their publication, with the exception of the 100% supervised baseline result (*), which we reproduced. Best scores for mean and standard deviation are bold, second best scores are underlined.

| Method | ACDC: DSC (%) ↑ Percentage of Labeled Patients | | | | MMWHS: DSC (%) ↑ Percentage of Labeled Patients | | | |
|---|---|---|---|---|---|---|---|---|
| | 2% | 4% | 16% | 100% | 10% | 20% | 80% | 100% |
| Supervised from [39] | 61.40 | 70.20 | 84.40 | **91.20** | 45.10 | 63.70 | 78.70 | 88.31 (*) |
| Noisy Student [38] | 63.20 | 73.70 | 83.60 | - | 59.30 | 68.50 | 78.00 | - |
| Mixup [31] | 69.50 | 78.50 | 86.30 | - | 56.10 | 69.00 | 79.60 | - |
| Self-Training [35] | 69.00 | 74.90 | 86.00 | - | 56.30 | 69.10 | 80.10 | - |
| LCLPL (inter) [39] | 75.90 | 83.10 | 88.30 | - | 57.20 | 71.90 | 81.10 | - |
| LCLPL (intra) [39] | 76.10 | 84.50 | 88.10 | - | 59.90 | 72.10 | 80.30 | - |
| $ST_{TC}$ (ours) | **84.53** | 87.90 | 89.66 | - | 64.63 | **86.09** | **88.64** | - |
| $ST_{TC}$-PL-$ST_{TC}$ (ours) | 78.00 | **88.36** | **90.01** | - | **67.54** | 85.81 | 88.56 | - |

**Table 4.** Results from comparison of our proposed method with Basak and Yin [40] on 2D ACDC dataset, using their exact evaluation setup. Three percentages of patients in the labeled set were investigated. Mean DSC performance measures of related methods were taken from their publication. We reproduced the supervised variants with our framework, as those were missing in [40]. Best scores for mean and standard deviation are bold, second best scores are underlined.

| Method | ACDC: DSC (%) ↑ Percentage Labeled | | |
|---|---|---|---|
| | 10% | 20% | 100% |
| Supervised (ours) | 86.54 | 88.63 | 91.92 |
| Supervised from [40] | - | - | **92.30** |
| Double-UA [27] | 83.30 | - | - |
| DTC [20] | 82.70 | 86.30 | - |

**Table 4.** *Cont.*

| Method | ACDC: DSC (%) ↑ Percentage Labeled | | |
|---|---|---|---|
| | **10%** | **20%** | **100%** |
| MC-Net [21] | 86.30 | 87.80 | - |
| MC-Net+ [22] | 87.10 | 88.50 | - |
| LCLPL [39] | 88.10 | 90.50 | - |
| PLGCL [40] | 89.10 | 91.20 | - |
| $ST_{TC}$ (ours) | 91.20 | **91.52** | - |
| $ST_{TC}$-PL-$ST_{TC}$ (ours) | **91.57** | 91.51 | - |

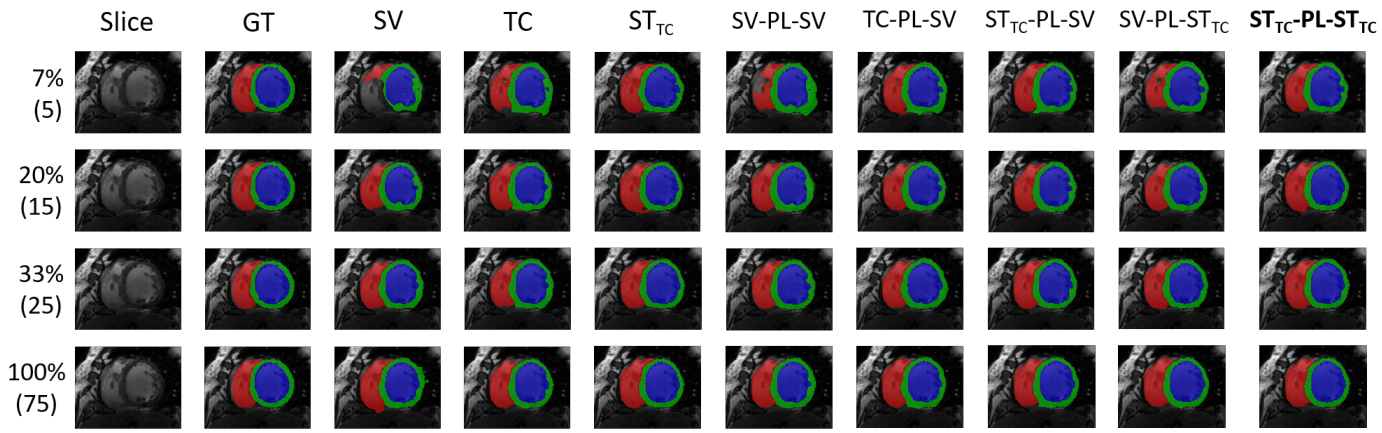



**Figure 6.** Exemplary qualitative results from ACDC using test subject ID 006. Rows refer to different percentages of labeled samples in the supervised set with the number of labeled patients given in brackets. While $ST_{TC}$ already achieves improvements over SV, even for solely 7% labeled images, the cascaded approaches, and especially $ST_{TC}$-PL-$ST_{TC}$, give overall best results.

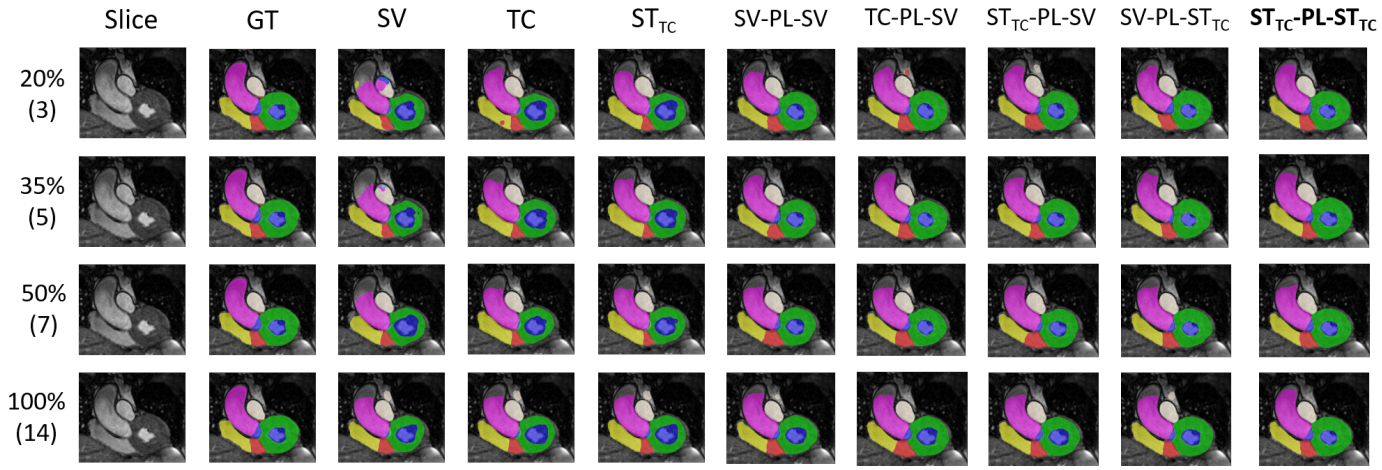



**Figure 7.** Exemplary qualitative results from MMWHS, using test subject ID 1001. Rows refer to different percentages of labeled samples in the supervised set with the number of labeled patients given in brackets. Again, $ST_{TC}$-PL-$ST_{TC}$ delivers very promising predictions even in the low labeled data regime.

### *5.1. Internal Evaluation*

Overall, the results of our evaluation shown in Tables 1 and 2 demonstrate that using unlabeled data via any baseline self-training method improves segmentation results. This is the case for all scenarios, where a restricted number of labeled patients is used, and also for fully supervised (100%) scenarios, indicating that adding unlabeled data is never detrimental and can thus always be considered. Importantly, the performance gains of self-

training compared with the entirely supervised variant become more prominent the smaller the size of the labeled set is, which highlights the effectiveness of SSL in the low-labeled data regime. We thus argue that the unsupervised losses can generally distill meaningful additional information from the unsupervised samples in cardiac MR segmentation. This aids in reducing tedious and costly labeling effort.

Out of the three baseline self-training methods, $ST_{TC}$ shows the largest performance increases over the supervised baseline, even giving overall best results in a few experiments. It seems to effectively combine the advantages of Transformation Consistency and Student–Teacher and is therefore studied in more detail in the various experiments when combined with pseudo-labeling methods. When introducing pseudo-labeling, additional performance gains can be observed, most strongly in the low-labeled data regime. While the pseudo-labeling baseline SV-PL-SV seems to have limitations due to the quality of initial pseudo-labels not being good enough, starting from $ST_{TC}$-based pseudo-labels it outperforms most simpler methods even when using a purely supervised training in the second round. However, in most experiments, supervised training in the second round can not fully compete with a cascaded self-training approach as proposed. We tested three cascaded Student-Teacher variants and found that the $ST_{TC}$-PL-$ST_{TC}$ approach shows the most promising results in terms of performance gains for ACDC and MMWS across all studied metrics and all label percentage setups. Except for a few still close results, $ST_{TC}$-PL-$ST_{TC}$ is either the best performing or second best performing method and shows low standard deviations among repetitions of the evaluations with different random seeds for training. Most notably, the boost between purely supervised and our proposed cascaded ST method in the lowest data regimes for ACDC and MMWHS is 10.17% and 6.72% in DSC, respectively. Our findings from the quantitative results are also confirmed when looking at predictions qualitatively. Both for ACDC (Figure 6) and MMWHS (Figure 7), the purely supervised approach (SV) shows segmentation errors in the low-labeled data regime. However, switching to $ST_{TC}$ gives great improvements, whereas including SV into Pseudo-Labeling is not always beneficial (see Columns 6 to 9, especially for ACDC). Qualitative results are consistently best when using our cascaded self-supervised approach $ST_{TC}$-PL-$ST_{TC}$. We conclude from our extensive experiments that it is always beneficial to train a cascade of ST networks when unlabeled data are available in the cardiac MRI segmentation setting.

### *5.2. Comparison to Literature*

Two recent works with state-of-the-art SSL methods, LCLPL [39] and PLGCL [40], were used to perform a comparison with our proposed $ST_{TC}$-PL-$ST_{TC}$ method. Chaitanya et al. [39] evaluated several SSL methods on MMWHS and ACDC and also assessed the performance in the most challenging data regime evaluated in this work, where solely one patient was used in the labeled dataset. For both ACDC and MMWHS (see Table 3), they showed that when trained solely supervised, their reduced labeled data results are far from the corresponding supervised baseline, which is using 100% of available labeled data. LCLPL is able to bridge this gap using SSL in all scenarios and coming close to supervised baselines for as little as 16% or 20% labeled samples, respectively. However, both our compared methods, the baseline $ST_{TC}$ and the proposed $ST_{TC}$-PL-$ST_{TC}$, outperform all related works drastically on both ACDC and MMWHS. Specifically, for 16% on ACDC and 20% on MMWHS, they perform on par with the upper bound supervised baseline, a behavior that was also seen in our comprehensive internal evaluation. In the most challenging 2% ACDC setting, we can see that our cascaded self-training, although outperforming all related work, is not competitive when compared with baseline Student–Teacher. We assume that when using solely one labeled patient in the first round, pseudo-labeling seems to overfit to this

sample, limiting its overall performance. However, with as little as 4% labeled subjects, the proposed cascaded self-supervised approach is already beneficial with a 3.86% improvement in DSC. The behavior for MMWHS indicates that there is no clear winner between baseline Student–Teacher and the cascaded version. We argue that in this setup, the very low number of actually used unlabeled samples (10) is not sufficient to significantly benefit from. In comparison, our internal evaluation showed that using the remaining 40 unlabeled samples of MMWHS lead to consistently better cascaded self-training results. However, in the very-low-data regime setup, $ST_{TC}$-PL-$ST_{TC}$ still outperforms the best related work by 7.64% DSC.

Our results in Table 4 demonstrate that our proposed method is also able to outperform the approaches in the ACDC-based evaluation setup that was used by Basak and Yin [40] to assess their PLGCL method. In the first reduced labeled patient setup, with seven subjects form the labeled set (10%), our $ST_{TC}$-PL-$ST_{TC}$ method performs the best compared with all presented methods, including also LCLPL [39]. The difference in DSC to the second best PLGCL method is $+2.47\%$. In the other reduced label setup, which includes 14 labeled subjects (20%), $ST_{TC}$ and $ST_{TC}$-PL-$ST_{TC}$ show almost identical results and again surpass all other presented methods, with a DSC difference to PLGCL of $+0.3\%$. Notably, the performance gap to the fully supervised baseline using a total of 70 patients again is nearly closed by our proposed cascaded self-training method in this experiment. We assume that the much larger number of unlabeled patients (70) is very beneficial in this experimental setup by [40], as opposed to the setup from [39].

### *5.3. Limitations*

While our experiments are comprehensive for the 2D and 3D cardiac imaging domain, our strategies are supposedly more generally applicable. Thus, our conclusions are necessarily limited to this domain and an extension to other domains, e.g., abdominal organ segmentation, might be valuable next steps to follow up on. Moreover, we currently do not consider whether self-supervised training might help in generalizing within dynamical scenarios, e.g., when training a model for one cardiac timepoint and applying it to a different timepoint, making this another interesting direction for future work.

Methodologically, from our comparison to recent related work, which makes use of a contrastive loss component, we can see that our baseline and cascaded approaches do not require such a component to achieve state-of-the-art results. Nevertheless, due to contrastive losses being widely used in computer vision and medical image analysis applications recently, it may be beneficial to additionally incorporate such a loss term into our unsupervised component. We intend to study such a combination in future work.

## 6. Conclusions

In this work, we comprehensively evaluated semi- and self-supervised learning strategies in the context of cardiac MR image segmentation. By studying transformation consistency, Student–Teacher, Pseudo-Labeling, and combinations thereof, we found a novel, highly promising combination by cascading two Student–Teacher SSL training rounds within a Pseudo-Labeling workflow. Our experiments on the 2D ACDC and 3D MMWHS training setup with reduced labeled datasets revealed that given a set of unlabeled samples which is often abundantly available, it is always beneficial to train segmentation networks in a self-supervised manner. Moreover, we experimentally demonstrated that even in very-low-labeled data regimes, we can improve upon supervised low data baselines (10.17% and 6.72% improvement in DSC for ACDC and MMWHS, respectively) but also upon recent state-of-the-art SSL techniques that use more sophisticated training strategies like contrastive learning (2.47% and 7.64% DSC improvement, respectively). Finally, we also

conclude that with our proposed cascaded Self-Supervision strategy, it is even possible to nearly close the performance gap to the fully supervised scenarios, where all available labeled samples are used during training.

**Author Contributions:** M.U., conceptualization, methodology, funding acquisition, visualization, writing—original draft; E.R., methodology, software, investigation; F.T., methodology, software, supervision, writing—review and editing; D.Š., conceptualization, methodology, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1748423.

**Data Availability Statement:** The data sets presented in this study are available publicly.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ACDC | Automated Cardiac Diagnosis Challenge |
| AO | Aorta |
| ASSD | Average Symmetric Surface Distance |
| CNN | Convolutional Neural Network |
| CT | Computed Tomography |
| CVD | Cardiovascular Diseases |
| DSC | Dice Similarity Coefficient |
| DTC | Dual Task Consistency |
| EMA | Exponential Moving Average |
| GDL | Generalized Dice Loss |
| LCLPL | Local Contrastive Loss with Pseudo-Labels |
| LA | Left Atrium |
| LV | Left Ventricle |
| MICCAI | Medical Image Computing and Computer Assisted Intervention |
| MMWHS | Multi-Modality Whole Heart Segmentation |
| MRI | Magnetic Resonance Imaging |
| MSE | Mean Squared Error |
| MYO | Myocardium |
| PA | Pulmonary Artery |
| PL | Pseudo-Labeling |
| PLGCL | Pseudo-Label Guided Contrastive Learning |
| RA | Right Atrium |
| RV | Right Ventricle |
| ST | Student–Teacher |
| SV | Supervised |
| TC | Transformation Consistency |
| SSL | Self-Supervised Learning |

## References

1. Roth, G.A.; Mensah, G.A.; Johnson, C.O.; Addolorato, G.; Ammirati, E.; Baddour, L.M.; Barengo, N.C.; Beaton, A.Z.; Benjamin, E.J.; Benziger, C.P.; et al. Global Burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 Study. *J. Am. Coll. Cardiol.* **2020**, *76*, 2982–3021. [CrossRef] [PubMed]
2. Wang, T.J. Assessing the role of circulating, genetic, and imaging biomarkers in cardiovascular risk prediction. *Circulation* **2011**, *123*, 551–565. [CrossRef] [PubMed]
3. Cawley, P.J.; Maki, J.H.; Otto, C.M. Cardiovascular magnetic resonance imaging for valvular heart disease: technique and validation. *Circulation* **2009**, *119*, 468–478. [CrossRef] [PubMed]

4. Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; Rueckert, D. Deep Learning for Cardiac Image Segmentation: A Review. *Front. Cardiovasc. Med.* **2020**, *7*, 25. [CrossRef]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
6. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
7. Zhuang, X.; Li, L.; Payer, C.; Stern, D.; Urschler, M.; Heinrich, M.P.; Oster, J.; Wang, C.; Smedby, O.; Bian, C.; et al. Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge. *Med. Image Anal.* **2019**, *58*, 101537. [CrossRef]
8. Zhuang, X.; Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Med. Image Anal.* **2016**, *31*, 77–87. [CrossRef]
9. Chapelle, O.; Schoelkopf, B.; Zien, A. *Semi-Supervised Learning*; The MIT Press: Cambridge, MA, USA, 2006.
10. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
11. Yang, X.; Song, Z.; King, I.; Xu, Z. A survey on deep semi-supervised learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 8934–8954. [CrossRef]
12. Neff, T.; Payer, C.; Štern, D.; Urschler, M. Generative Adversarial Network based Synthesis for Supervised Medical Image Segmentation. In Proceedings of the OAGM&ARW Joint Workshop 2017: Vision, Automation and Robotics, Vienna, Austria, 10–12 May 2017; pp. 140–145. [CrossRef]
13. Hadzic, A.; Bogensperger, L.; Joham, S.J.; Urschler, M. Synthetic Augmentation for Anatomical Landmark Localization Using DDPMs. In Proceedings of the Simulation and Synthesis in Medical Imaging (SASHIMI 2024), Marrakesh, Morocco, 10 October, 2025; Volume 15187, Lecture Notes in Computer Science, pp. 1–12. [CrossRef]
14. Jiao, R.; Zhang, Y.; Ding, L.; Xue, B.; Zhang, J.; Cai, R.; Jin, C. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Comput. Biol. Med.* **2024**, *169*, 107840. [CrossRef]
15. Bortsova, G.; Dubost, F.; Hogeweg, L.; Katramados, I.; de Bruijne, M. Semi-supervised Medical Image Segmentation via Learning Consistency Under Transformations. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Shenzhen, China, 13–17 October 2019; Springer International Publishing: Cham, Switzerland, 2019; pp. 810–818. [CrossRef]
16. Li, X.; Yu, L.; Chen, H.; Fu, C.W.; Xing, L.; Heng, P.A. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 523–534. [CrossRef] [PubMed]
17. DeVries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552. [CrossRef]
18. Yun, S.; Han, D.; Chun, S.; Oh, S.J.; Yoo, Y.; Choe, J. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
19. French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; Finlayson, G. Semi-supervised semantic segmentation needs strong, varied perturbations. In Proceedings of the British Machine Vision Conference (BMVC), Virtual Event, 7–10 September 2020; pp. 1–14. [CrossRef]
20. Luo, X.; Chen, J.; Song, T.; Wang, G. Semi-supervised medical image segmentation through dual-task consistency. *Proc. Conf. AAAI Artif. Intell.* **2021**, *35*, 8801–8809. [CrossRef]
21. Wu, Y.; Xu, M.; Ge, Z.; Cai, J.; Zhang, L. Semi-supervised left atrium segmentation with mutual consistency training. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Strasbourg, France, 27 September–1 October 2021; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2021; pp. 297–306. [CrossRef]
22. Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; Cai, J. Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **2022**, *81*, 102530. [CrossRef] [PubMed]
23. Huang, H.; Chen, Z.; Chen, C.; Lu, M.; Zou, Y. Complementary consistency semi-supervised learning for 3D left atrial image segmentation. *Comput. Biol. Med.* **2023**, *165*, 107368. [CrossRef]
24. Laine, S.; Aila, T. Temporal Ensembling for Semi-Supervised Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
25. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

26. 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 Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019, Shenzhen, China, 13–17 October 2019; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 605–613. [CrossRef]
27. Wang, Y.; Zhang, Y.; Tian, J.; Zhong, C.; Shi, Z.; Zhang, Y.; He, Z. Double-uncertainty weighted method for semi-supervised learning. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, Lima, Peru, 4–8 October 2020; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 542–551. [CrossRef]
28. Liu, L.; Tan, R.T. Certainty driven consistency loss on multi-teacher networks for semi-supervised learning. *Pattern Recognit.* **2021**, *120*, 108140. [CrossRef]
29. Huang, W.; Chen, C.; Xiong, Z.; Zhang, Y.; Chen, X.; Sun, X.; Wu, F. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Trans. Med. Imaging* **2022**, *41*, 3016–3028. [CrossRef]
30. Lei, T.; Zhang, D.; Du, X.; Wang, X.; Wan, Y.; Nandi, A.K. Semi-Supervised Medical Image Segmentation Using Adversarial Consistency Learning and Dynamic Convolution Network. *IEEE Trans. Med. Imaging* **2023**, *42*, 1265–1277. [CrossRef]
31. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
32. Basak, H.; Bhattacharya, R.; Hussain, R.; Chatterjee, A. An exceedingly simple consistency regularization method for semi-supervised medical image segmentation. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4. [CrossRef]
33. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl. Inf. Syst.* **2015**, *42*, 245–284. [CrossRef]
34. Beyer, L.; Zhai, X.; Oliver, A.; Kolesnikov, A. S4L: Self-supervised semi-supervised learning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
35. Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P.M.; Rueckert, D. Semi-supervised learning for network-based cardiac MR image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2017, Quebec City, QC, Canada, 11–13 September 2017; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; pp. 253–260. [CrossRef]
36. Fan, D.P.; Zhou, T.; Ji, G.P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans. Med. Imaging* **2020**, *39*, 2626–2637. [CrossRef]
37. Li, Y.; Chen, J.; Xie, X.; Ma, K.; Zheng, Y. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, Lima, Peru, 4–8 October 2020; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 614–623. [CrossRef]
38. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-training with noisy student improves ImageNet classification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
39. Chaitanya, K.; Erdil, E.; Karani, N.; Konukoglu, E. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Med. Image Anal.* **2023**, *87*, 102792. [CrossRef]
40. Basak, H.; Yin, Z. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: Piscataway, NJ, USA, 2023. [CrossRef]
41. Payer, C.; Štern, D.; Bischof, H.; Urschler, M. Multi-label Whole Heart Segmentation Using CNNs and Anatomical Label Configurations. In Proceedings of the Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges, Quebec City, QC, Canada, 10–14 September 2018; Springer International Publishing: Cham, Switzerland, 2018; pp. 190–198. [CrossRef]
42. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In Proceedings of the Deep Learning in Medical Image Analysis—DLMIA 2017, Quebec City, QC, Canada, 14 September 2017; Springer: Cham, Switzerland, 2017; Volume 10553, Lecture Notes in Computer Science, pp. 240–248. [CrossRef]
43. Chen, X.; He, K. Exploring simple Siamese representation learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021. [CrossRef]

44. Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.A.; Cetin, I.; Lekadir, K.; Camara, O.; Gonzalez Ballester, M.A.; et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging* **2018**, *37*, 2514–2525. [CrossRef]
45. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
46. Thaler, F.; Gsell, M.A.F.; Plank, G.; Urschler, M. CaRe-CNN: Cascading Refinement CNN for Myocardial Infarct Segmentation with Microvascular Obstructions. In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024)—Volume 3: VISAPP, Rome, Italy, 27–29 February 2024; pp. 53–64. [CrossRef]
47. Thaler, F.; Stern, D.; Plank, G.; Urschler, M. LA-CaRe-CNN: Cascading Refinement CNN for Left Atrial Scar Segmentation. In Proceedings of the MICCAI Challenge on Comprehensive Analysis and Computing of Real-World Medical Images, CARE 2024, Marrakesh, Morocco, 6–10 October 2024; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15548, pp. 180–191. [CrossRef]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1026–1034.
49. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
50. Payer, C.; Stern, D.; Bischof, H.; Urschler, M. Coarse to Fine Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)—Volume 5: VISAPP, Valletta, Malta, 27–29 February 2020; pp. 124–133. [CrossRef]
51. Luo, X.; Wang, G.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Metaxas, D.N.; Zhang, S. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med. Image Anal.* **2022**, *80*, 102517. [CrossRef]