Dynamic contrastive learning guided by class confidence and confusion degree for medical image segmentation

Jingkun Chen[a,b], Changrui Chen[b], Wenjian Huang[a], Jianguo Zhang[a,d,*], Kurt Debattista[b,*], Jungong Han[b,c,*]

[a] *Department of computer science and engineering, Southern University of Science and Technology, Shenzhen, 518055, China*
[b] *WMG Visualization, University of Warwick, Coventry, CV4 7AL, United Kingdom*
[c] *Department of Computer Science, University of Sheffield, Sheffield, S10 2TN, United Kingdom*
[d] *Peng Cheng Lab, Shenzhen, China, 518000, China*

Abstract

This work proposes an intra-Class-confidence and inter-Class-confusion guided Dynamic Contrastive (CCDC) learning framework for medical image segmentation. A core contribution is to dynamically select the most *expressive pixels* to build positive and negative pairs for contrastive learning at different training phases. For the positive pairs, dynamically adaptive sampling strategies are introduced for sampling different sets of pixels based on their hardness (namely the easiest, easy, and hard pixels). For the negative pairs, to efficiently learn from the classes with high confusion degree w.r.t a query class (i.e., a class containing the query pixels), a new *hard class* mining strategy is presented. Furthermore, pixel-level representations are extended to the neighbourhood region to leverage the spatial consistency of adjacent pixels. Extensive experiments on the three public datasets demonstrate that the proposed method significantly surpasses the state-of-the-art. The code is publicly available at https://github.com/jingkunchen/ccdc.

Keywords: Class Confusion Degree, Dynamic Contrastive Learning,

[☆] This work is supported in part by National Key Research and Development Program of China (2021YFF1200800) and National Natural Science Foundation of China (Grant No. 62276121)

[*] Corresponding authors

1. Introduction

Deep learning based medical image segmentation algorithms are capable of providing clinical guidance under the right conditions, such as surgery planning and post-surgery monitoring. However, training a successful deep-learning-based segmentator requires a large amount of pixel-level annotation data, which is always rare and expensive, especially for 3D images. It remains a challenge to train a satisfactory deep learning-based segmentation algorithm with limited annotated data.

In recent years, contrastive self-supervised representation learning is widely adopted to reduce the use of labelled data, which turns out to be successful in solving *image-level* tasks, such as classification [1] and image translation [2], and *instance-level* tasks, such as detection [3]. In contrastive learning, constructions of positive (similar) pairs and negative (dissimilar) pairs [4] are essential. This is not trivial, especially when dealing with a large number of pixels, e.g., segmentation task, as naive contrastive learning that exhaustedly combines all pixels becomes computationally prohibitive. Therefore, sampling a portion of pixels for contrastive learning in segmentation task is usually adopted as an alternative [5].

To build positive pairs in contrastive learning, strategies of sampling pixels from images for segmentation task, such as random sampling [6] and active hard sampling [5], suffer from several limitations. More concretely, random sampling [6] treats each pair equally without considering the effectivenesses of the selected pixels. Often, a large pixel sampling ratio will inevitably introduce redundant information, as not all pairs are needed at the same time. For example, easy pixels cannot contribute to model optimisation when the performance of the model is high. Alternatively, active hard sampling strategies [5] focus only on hard pixels but exclude easily distinguished samples during all training phases. This may lead to hindered convergence due to the sharp gradient fluctuations when feeding hard pixels to a model with unstable/or poor performance [7]. Compared to using hard pixels, the *non-hard* pixels are more conducive to training when the performance of a model is poor.

On the other hand, for negatives sampling, pixels sampled from an easy class tend to contribute less to the optimisation of a model, especially in the case of class imbalance (easily distinguishable classes contain more pixels, while hard classes contain fewer pixels) [8]. The sampled negative pixels are more useful for learning discriminative feature only when the pixels from hard classes are sampled frequently.

Thus, two to-be-solved problems appear when applying pixel-level contrastive learning in segmentation task: Firstly, as the number of pixels in a dataset is huge, efficiently sampling representative pixels is the main problem to be solved. Secondly, the needs of pixels of a model will change at different phases. It is more reasonable to use appropriate sampling strategies during different training phases. Solving these questions can provide insights to promote the design of an efficient pixel-level contrastive learning sampling strategy. We first investigate the first question by exploring the existing pixel-level contrastive learning methods, which sample pixels from a class to construct positive pairs (z and z^{+}) and different classes to construct negative pairs (z and z^{-}). However, most existing efforts treat all pixels equally and do not consider the hardness of the pixels. In an image containing regions from different classes, the center pixels in one class-specific region tend to be easily distinguishable than the border pixels, as border pixels could be easily confused with neighbouring pixels belonging to other classes. Thus, to separate different regions within a class and treat the pixels in these regions differently, it is necessary to divide the pixels in a class into different groups according to their hardness and each group should be treated differently. For the second question, most of the works in the literature use the same sampling strategy at the different training phases. However, if there are significant changes in the performance of the model, a fixed sampling strategy cannot always sample representative pixels to contribute to the model. For example, when the performance of the model is already good enough, extracting

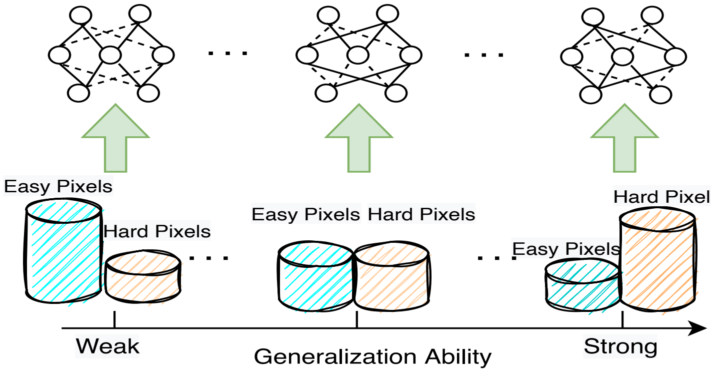

Figure 1: Our intra-class confidence degree guided dynamic contrastive learning aims to select different sampling strategies to build positive pairs at different training phases. The dashed lines with arrows denote the data flow; the solid lines are used to indicate the construction of the positive pairs. The height of the cylinder represents the number of easiest, easy and hard pixels in Class A in the current training phase. The length of the rectangle indicates the number of samples. When the intra-class confidence degree of a model is low, a small sampling ratio to the hard pixels and a much larger sampling ratio to the easy pixels is applied for training. When the intra-class confidence degree of a model is strong, the sampling strategy gradually moves to the opposite ratio for them.

a large number of easy pixels does not contribute much to the optimisation. On the contrary, taking more hard pixels when the performance of a model is low will produce a large gradient, which may prevent the model from converging efficiently.

Based on the above discussion, we first propose to dynamically change the sampling strategy at different training stages to select the most expressive pixels. Based on this idea, a *dynamic intra-class confidence degree* is introduced to divide the pixels into three groups: query pixels (z), intra-class core (z^{+}) and inter-class negative pixels (z^{-}). Query pixels (z) and the intra-class core (z^{+}) are used to construct positive pairs, and query pixels (z) and inter-class negative pixels (z^{-}) are used to construct negative pairs. In the following part of this paper, we use the pixels of a class A as an example to illustrate our contrastive strategy. For sampling queries and core, as shown in Fig. 1, several adaptive sampling strategies based on variations in model performance (low, medium, and high confidence degree for Class A) are introduced to select pixels. Specifically, for all phases, the *easiest* distinguishable pixels of class A can be considered as the core (z^+). The representations generated by the core (z^+) can be used as a prototype to boost other query pixels (z) in the same class for class-specific representation learning [9]. When the overall classification accuracy of the pixels in class A is low, the *easier* pixels in class A should be sampled more as queries for optimisation to reduce learning difficulty and improve overall performance [10]. When the accuracy of the pixels in class A is medium, the sampling ratio of the *easier* query pixels should be reduced and the *harder* ones should be further considered, i.e. starting easy and gradually presenting more complex concepts [11] can obtain better learning effects. If accuracy is high, more attention should be paid to the *hard* pixels to improve the hardest problem-solving skills. It is worth noting that our approach is different from the easy-to-hard sampling strategy [12], as we dynamically measure the performance of the model on each class at different phases and adapt accordingly. The proposed method can dynamically change the sampling strategy to sample the most representative pixels to train a more robust model (verified in Sect. 4.3.2) and avoid the risk of the model crashing due to a catastrophic incident (the great variance of gradients) by the sampled hard pixels [7].

Generally, pixels sampled from more confusing classes as negatives can improve the efficiency of network optimisation, as hard classes pixels allow the model to focus on distinguishable details [13]. Since pixels from two adjacent classes or two similar classes are more difficult to distinguish, and pixels from two distant classes or dissimilar classes are often easier to distinguish, it is possible to focus on relatively difficult classes to get better performance. For example, as shown in Fig. 2, the right and left ventricles are two organs that are separated by the myocardium and not adjacent to each other. For the right and left ventricles, because of their different structures, colour and intensity distribution, the right ventricle can be easily separated from the left ventricle. But for the myocardium class, the neighbouring pixels in the myocardium class greatly influence the ventricle classes in adjacent regions. Based on this observation,

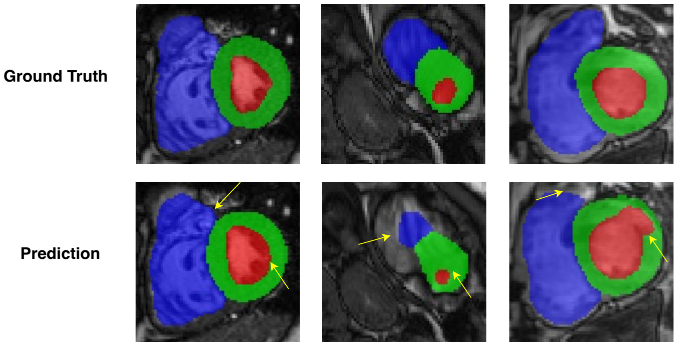

Figure 2: Cardiac images with ground truth masks, the left and right ventricles (Blue and Red regions) are two organs that are separated by the myocardium (Green region) and not adjacent to each other. Pixels in the ventricles class are often misclassified as myocardium class because pixels on the boundary are easily affected by other classes.

negative pixel sampling strategies should be designed *differently* from query sampling strategies which need to choose easy pixels when the performance is low and hard pixels when the performance is high. The proposed method helps determine which class is more confusing. The *inter-class-confusion degree* is introduced to measure the *inter-class* difficulty from one class to other classes at different phases. The inter-class difficulty can be used to control the number of sampled pixels from different negative classes.

Finally, in the segmentation task, local relationship is crutial for determining the boundary of two classes. A representation at the positive is definitely influenced by its neighbouring pixels [14]. We enhance the pixel-level representation by combining the CNN features of individual pixels with their neighbouring pixels. Experiments show that our region-based contrastive learning method is more effective than using the single pixel-level one.

In this paper, a new loss, intra-Class-confidence and inter-Class-confusion guided Dynamic Contrastive (CCDC) loss, is introduced to dynamically change the pixels sampling strategies at different training phases. We use the state-of-the-art nnU-Net [15] framework as the backbone to evaluate our approach on three datasets, the MS-CMRSeg Dataset [16], the ACDC Challenge Dataset [17] and CHAOS (combined (CT-MR) Healthy Abdominal Organ Segmentation) Dataset [18]. The proposed CCDC learning outperforms the state-of-the-art methods. The contributions of this paper are summarized below.

- We design a *intra-class-confidence and inter-class-confusion* guided dynamic contrastive framework and introduce a new loss, CCDC loss, for medical image segmentation. This is the first attempt to dynamically change the sampling strategy to select the most *expressive pixels* to construct positive and negative pairs based on the current circumstance of each class.
- A *intra-class-confidence degree* is introduced to measure the performance of a model on each class. A dynamically adaptive intra-class sampling strategy based on the intra-class-confidence-degree is used to separate the easiest, easy, and hard pixels for intra-class similarity learning.
- To learn the class differences, a *inter-class-confusion degree* is proposed to dynamically adjust the sampling ratio of the positive and negative classes, and prioritize sample pixels from the hard classes.
- We extend pixel-level contrastive learning to the region-level by considering the spatial dependency of neighboring pixels, which turns out to be more suitable for the segmentation task.
- The results of three public datasets show that the proposed method is more effective than other pixel-level contrastive learning methods, which is highly relevant in the field of semantic segmentation, especially in medical imaging where accurate segmentation is crucial for diagnosis and treatment.

2. Related Work

Although traditional computer vision techniques have proven to be effective in extracting representations from image data, the rise of Convolutional Neural

Networks (CNNs) has challenged their dominance and established a new benchmark for representation learning. The ability of CNNs to learn hierarchical representations from raw data, incorporating both local and global context, has been demonstrated through numerous studies and outperforms traditional computer vision methods in many tasks. In recent years, CNN-based approaches such as FCN [19], U-Net [20] and other related models [21] have gained significant popularity for medical image segmentation across various fields, including tasks such as lung nodules segmentation [22], hippocampus segmentation [23], ventricle and myocardium segmentation [24] and MRI-CT whole heart segmentation [25]. Furthermore, the application of Transformers in medical image segmentation has been on the rise as well, with notable models like EffTrans[26] and nnFormer [27]. For CNN- and transformer-based backbones, most segmentation models use pixel-level cross-entropy loss and Sorensen-Dice coefficient loss to train the networks. However, none of these loss functions takes explicitly into account the relationship between different pixels (whether the pixels come from the same classes), which can be fundamental for a segmentation task.

2.1. Pixel-level Contrastive learning

Contrastive learning is a form of comparative learning that involves forming positive pairs and negative pairs to learn the common representation between similar instances and discriminative representations between non-similar instances [28]. For classification problems, contrastive learning is generally performed on image-level representations. Positive samples are sampled from homologous images to ensure similarity, while negative samples are usually derived from non-homologous images. Generally, negative mining strategies in contrastive learning are widely used for image-level classification tasks to reduce computational effort and improve efficiency [29].

In the segmentation task, to exploit the potential of the labelled images, contrastive learning helps capture detailed information to consider the elaborate pixel-level representations rather than the whole image-level representation [30]. Although there have been some works on pixel-level contrastive learning, these methods have the following limitations:

(1) Sampling strategies such as *random sampling, uniform sampling* , are usually unstable and may result in missing necessary samples or taking unnecessary samples. *Hard pixel mining* strategies [6] focus on the hardest pixels and ignore the contribution of non-hard pixels. When the performance of a model is low, *non-hard* pixels should be given priority over the hardest ones to avoid the risk of model divergence due to the great variance of gradients by a large loss [7]. *Balancing weight of the easy and hard pixels*, such as Focal loss [31], gives less weight to easy pixels and gives more weight to hard misclassified pixels. However, It tends to produce vanishing gradients during backpropagation, penalizes negative classes in reverse, or employ non-optimal loss weights between classes [32].

(2) In the segmentation task, recent works ignore the spatial dependency between neighbouring pixels. A pixel-level representation has a strong spatial dependency on its neighbouring pixels. The pixels' spatial dependency in segmentation tasks is crucial to delineating the boundary. Treating all pixels individually would lose the spatial information between the pixels [14].

2.2. Dynamic Learning

For dynamic learning, the traditional approach is to learn from easy to hard, designed to mimic the meaningful learning strategies of human learning [33]. The basic idea of easy to hard learning is to initially train the model with simple data and then gradually feed more difficult samples to obtain a better generalization [12]. This strategy has shown potential for image-level classification [34] and object-level detection [35]. However, there are few studies on easy-to-hard strategies for pixel-level contrastive learning. Meanwhile, directly using the easy-to-hard learning strategy for pixel-level contrastive learning still suffers from limitations due to pixels sampling risk; a model with low performance caused by insufficient training with easy pixels may make model more unstable or even cause model to crash when hard pixel oversampling. Since learning with hard samples in a weak model introduces a larger uncertainty [7], sampling from easy to hard can reduce the risk of model collapse and build a robust model. However, continuing to sample harder pixels may be counterproductive when the model's performance decreases. Compared to the easy-to-hard strategy, as shown in Sect. 4.3.2, our adjusting sampling strategy dynamically samples different difficulty pixels based on the current performance of the model and can perform well in representation learning.

3. Methodology

The quality of the pixel-level representations is crucial to the performance of pixel-level segmentation. In this study, to improve the quality of the pixel-level representations, a dynamically pixel-level contrastive learning strategy, termed intra-class-confidence and inter-class-confusion guided dynamic contrastive learning, is designed to learn pixel-level representations with stronger intra-class similarity and inter-class distinguishability.

As shown in Fig. 3, we employ nnU-Net [15] as the backbone, the network is responsible for extracting feature maps from the input image, and the feature maps are denoted by h_i in the figure, followed by two heads: (1) one segmentation head with a Softmax function, the segmentation head is used for predicting the class label of each pixel and (2) one projection head is then applied on h_i to project them into a feature space where contrastive learning is performed, the feature maps after the projection head are denoted by z_i in the figure. The cross-entropy loss ($\mathcal{L}_{CE}$) and DICE loss ($\mathcal{L}_{DICE}$) are applied to the segmentation head branch. In the projection head, we introduce our proposed dynamic learning strategy for contrastive learning.

3.1. Pixel Grouping In Contrastive Learning

To conduct pixel-level contrastive learning, we first divide all pixels of images in a batch into three groups: query pixels (z), and intra-class core (z^+, which is used to build a class-specific prototype) and inter-class negative pixels (z^-, which are used for learning the class-distinguished details). In our contrastive learning framework, query pixels (z), and intra-class core (z^+) are used

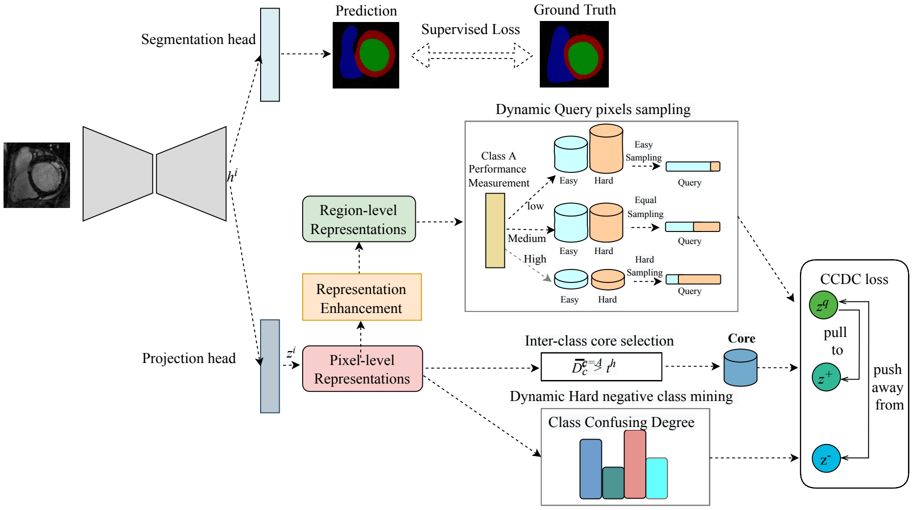

Figure 3: Model overview. Two modules are designed to enhance the feature learning of the backbone of the auto-encoder. One is for supervised learning against the ground truth and the other is for intra-class confidence degree and inter-class-confusion degree guided dynamic contrastive learning, with a projection head mapping the feature representations to the space where contrastive loss is applied.

to construct positive pairs, and query pixels (z) and inter-class negative pixels (z^-) are used to construct negative pairs.

3.2. Confidence Class-Specific Core

To sample pixels to build positive pairs for contrastive learning, we start by defining the *intra-class confidence degree*, which will be used to measure the performance of a model at different training phases. We first use the class-specific ground-truth mask to split all the pixels into different classes according to their ground truth class. In the following part of this section, we use the pixels of the class A as an example to illustrate our dynamic contrastive sampling strategy, where A is the ground truth class. For i-th pixel in a class, we denote the predicted probability of this class as $p_i^c = P(y_i = c|x_i)$, where c is the predicted class, and, y_i and x_i are the predicted label and pixel i in image x respectively. When $c = A$, we define a *intra-class confidence degree* by aggregating the probabilities $p_i^{c=A}$ for all pixels with respect to class A:

$$\overline{D}_{A}^{c=A} = avg(\{p_i^{c=A} | i \in \mathcal{I}_{\mathcal{A}}\}), \tag{1}$$

where $\mathcal{I}_A$ denotes the set of indices of all pixels in class A and $p_i^{c=A}$ is the probability pixel i predicted to belong to class A and avg is the average operation.

As shown in Fig. 1, we select the most representative pixels in class A as the core by intra-class confidence degree above a high threshold, as these pixels are the easiest to distinguish. The core is designed to be a class-specific prototype of the class A which can not only be used to distinguish this class from other classes but also can boost other intra-class pixels' learning. The core for the class A is defined as the *fused representation* produced by averaging these high confidence pixels' representations with predicted probabilities being larger than a threshold: $p_i^{c=A} > t^h$. Let z_i be the representation of pixel i. The core z^{A+} of the class A is calculated as

$$z^{A+} = \frac{\sum_i^{\mathcal{I}_\mathcal{A}} \alpha_i z_i}{\sum_i^{\mathcal{I}_\mathcal{A}} \alpha_i}, where \ \alpha_i = \mathbb{1}(p_i^{c=A} > t^h), \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

3.3. Intra-class Query Pixels Sampling Strategy

Instead of sampling only easy and/or hard queries [5], firstly, a dynamic sampling strategy that uses an *adaptive threshold* according to the model's performance on class A is introduced to split all the pixels into an easy query set and a hard query set. Then, *dynamically changed sampling ratios of* the easy and hard queries involved in training following the principle of sampling the most representative pixels. As is shown in Fig. 1, we divide all $p_i^{c=A}$-sorted representations of the class A (the pixel-wise probability of class A, sorted in descending order according to the adaptive threshold derived from the inter-class confidence degree of class A) into easy and hard subsets via the *adaptive threshold* derived from the inter-class confidence degree of the class A itself. It is worth noting that not only the *easy-hard sampling ratio* but also the *threshold* for distinguishing between easy and hard pixels are both *dynamically changed* at different training phases.

3.3.1. Adaptive Threshold for Query Pixels Sampling

We introduce an adaptive threshold for query pixels sampling. The adaptive threshold is formulated in Eq. 3.

$$T^A = 1 - \overline{D}_A^{c=A}. \tag{3}$$

Here, $\overline{D}_A^{c=A}$ is the intra-class-confidence degree of class A. For example, when $\overline{D}_A^{c=A}$ is 0.3, the representations with $p_i^{c=A} < (1-0.3)$ are marked as hard queries while $p_i^{c=A} > (1-0.3)$ are categorized as easy queries. When performance in class A is poor, a high threshold is preferred so that pixels above this high threshold can be grouped into the easy subset, thus sampling pixels from this easy subset to train the model with low performance. When performance is high, a lower threshold is chosen so that the harder pixels can be placed in the hard subset, thus the harder pixels can be used for a model to improve the ability to solve hard problems.

3.3.2. Dynamically Adaptive Sampling Ratios of the Easy and Hard Subsets

The $\overline{D}_A^{c=A}$ also assumes the responsibility of adjusting the ratio of easy and hard sample queries. The sampling ratios of easy/hard are set as $(1 - \overline{D}_A^{c=A})/\overline{D}_A^{c=A}$, which aims to sample more easier pixels when the performance of Class A is low and sampling harder pixels when the performance of class A is high. For example, when $\overline{D}_A^{c=A}$ is 0.3, that means the predicted ability of class A is weak. During this training phase, easy pixels are more acceptable for optimisation. Thus, the sampling ratios of the easy and hard queries should be 0.7 and 0.3. On the contrary, at a phase where the model is highly generalizable, the well-trained network should focus on the hard samples to further improve model performance.

3.4. Negative Pixels Sampling Strategy

Sampling negative pixels for contrastive learning will help the model learn the class-differentiable details. Classes that are different from the query class are treated as negative classes. The inter-class-confusing degree $\overline{D}_A^{c\neq A}$ in Eq. 4 can represent the misclassification probability of all the negative classes.

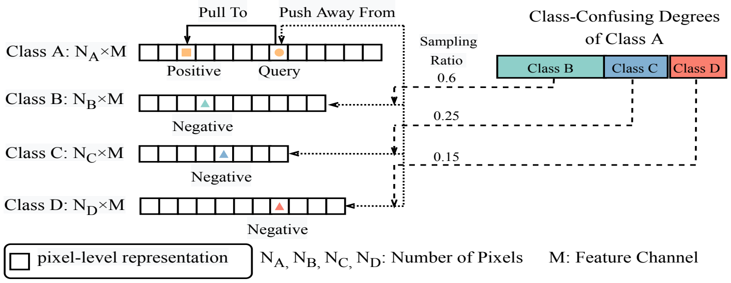

Figure 4: Hard negative class sampling strategy. The representations of all pixels in the image are divided into four groups by class labels, and each group contains the representations of all pixels within a class. The query pixel (a circle) in class A (yellow) is pulled to the positive core (prototype, represented as a square) generated by the easiest pixels in A and pushed away from negatives (triangles), namely the pixels of the other classes (Green, Blue and Red). The sampling rate of samples from the negative classes is determined by the class confusing degree of the query class A. The difference in the lengths of the rectangles of (Class B, C and D) reflects the degree of confusion between class A and other classes.

$$\overline{D}_A^{c\neq A} = avg(\{p_i^{c\neq A} | i \in \mathcal{I}_\mathcal{A}\}). \tag{4}$$

To separate the hardest classes from class A, we identify hard negative classes based on *inter-class-confusing degrees* $\overline{D}_A^{c\neq A}$. When selecting negative samples, the negative class with the highest confusion with respect to query class A should be given a higher sampling ratio to further discriminate queries from this negative class. While a class with a small negative class confusion degree is easier to distinguish and thus does not need to sample many negative samples. We use the ratio $[\overline{D}_A^1 : \overline{D}_A^2 :... :\overline{D}_A^c]$ where $c \neq A$, to build negative class sampling ratios. As shown in Fig. 4, Class A is misclassified as Class B with a higher probability of misclassification at 0.6 than Class C (0.25) and Class D (0.15). When a pixel of Class A is used as the query class, the other classes are regarded as negative classes, and the negative samples of the different negative classes should follow the confusing degrees of the class [0.6: 0.25: 0.15] to sample different numbers of pixels in each negative class. If the inter-class-confusing degrees of Class A with the other classes changes as the model's performance

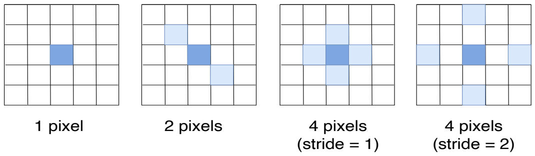

Figure 5: Region spatial layout. The pixel and its neighbors together represent the regional representation of the pixel, and the neighbors include horizontal, vertical and diagonal pixels.

changes, the sampling ratios of the different classes should also be dynamically adjusted.

3.5. Contrastive Enhancement with Spatial Prior

To incorporate spatial dependency into query optimisation by a *spatial enhancement strategy*, the pixel-level queries are enhanced with their neighbours to build region-level queries for contrastive learning. As shown in Fig. 5, we generalize representations of each query pixel with the features of its neighbouring pixels to build region-level representations z^q in the segmentation task, as pixels are spatially correlated.

3.6. Class Confidence and Confusion Guided Dynamic Pixels Sampling Strategy for Contrastive Learning

We use the proposed dynamic sampling strategies in the above sections to sample core, query and negative pixels to construct positive and negative pairs for contrastive learning, intending to select the pixels that are most beneficial for improving the performance. In this section, We will describe in detail how these sampled pixels are used in our contrastive learning framework.

At different training phases, the intra-class confidence degree $\overline{D}_A^{c=A}$ is used to reflect the performance of class A at each training phase and the inter-class confusing degree $\overline{D}_A^{c\neq A}$ is used to reflect misclassification degree of the pixels in class A to the other classes. In the following part of this section, we use θ^+ as the intra-class-confidence degree and θ^- as the inter-class-confusing degree to dynamically change the contrastive sampling strategy. We aim to use different pixel sampling strategies to train a model with different performances. Firstly, we define a joint probability distribution of the easy pixels z_e^q and the hard pixels z_h^q which belong to class A over classes C:

$$P(A|\{z_e^q, \lambda z_e^h\}) = \frac{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+})}{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+}) + \sum_{c\neq A}^{C} exp(-s^{\theta^-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-})}, \tag{5}$$

with $s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+} = min\{< \{z_e^q, \lambda z_h^q\}, z_A^+ >\}$. $s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+} \in [-1, 1]$ is used to measure the similarity between the query pixels and the core (easiest pixels) of class A. θ^+ is the intra-class confidence degree to control the easy-hard subsets dividing threshold and the sampling ratio λ of intra-class easy and hard query pixels. $s^{\theta^-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-}$ is the query-negative distance. θ^- is the inter-class-confusing degree to control the sampling ratio β of negative pixels of different classes. The CCDC loss of class A for mixed easy and hard queries with a confidence core and hard priority classes negative pixels can be expressed as:

$$\begin{aligned}\mathcal{L}^A_{CCDC} &= -logP(A|\{z_e^q, \lambda z_e^h\}) \\ &= -log\frac{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+})}{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+}) + \sum_{c\neq A}^{C} exp(-s^{\theta^-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-})}.\end{aligned} \tag{6}$$

Our designed CCDC loss for all the classes can be formulated as:

$$\mathcal{L}_{CCDC} = \frac{1}{C}\sum_{a\in C} -log\frac{exp(-s^{\theta^+}_{\{i_e,\lambda i_h\},a})}{exp(-s^{\theta^+}_{\{i_e,\lambda i_h\},a}) + \sum_{c\neq a}^{C} exp(-s^{\theta^-}_{i,\beta a})}. \tag{7}$$

The designed CCDC loss measures similarity via the dot product of region-level representations. It is worth mentioning that our CCDC loss is different from the other contrastive loss [5, 6, 36] since the queries, core and negatives in our CCDC, are all dynamically adjusted by different sampling strategies at different training phases. Moreover, our CCDC loss also differs from pixel-only sampling in that we extend the representation to a regional level that contains spatial information. The overall loss is the sum of the cross-entropy loss, the DICE loss, and the CCDC loss in the following way:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{DICE} + \mathcal{L}_{CCDC}. \quad (8)$$

4. Experiment

4.1. Experimental Setup

Dataset. We use three publicly available datasets to evaluate our method. (1) The MS-CMRSeg dataset [17] contains 45 3D cases. Each case is composed of 10-16 2D slices, resulting in a total of 686 slices. Following the setting of the challenge, the LGE modality is applied to our experiments. 10% of the data (5 3D scans) is used for training, and 90% of the data (40 3D scans) is used for testing. (2) CHAOS (Combined (CT-MR) Healthy Abdominal Organ Segmentation) Dataset [18] provides 60 3D DICOM data from 20 patients [T1-DUAL in phase (20 cases), T1-DUAL out phase (20 cases) and T2-SPIR (20 cases)] with four organs ground truth masks: liver, right kidneys (RK), left kidneys (LK) and spleen. The dataset was obtained from a 1.5T Philips MRI, which produces 12-bit DICOM images. The number of slices is between 26 and 50, resulting in a total of 1917 slices. The resolution is approximately 224 × 320. 10% of the data (6 3D scans from 2 patients) is used for training and the remaining 90% of the data (54 3D scans from 18 patients) is used for testing. There is no overlap in patient IDs between the training and testing sets. (3) The ACDC Challenge Dataset [16] consists of 100 patients (200 3D scans), with each case comprising 6-16 2D slices, resulting in a total of 1658 slices. The dataset was acquired using 1.5T and 3T scanners and is accompanied by expert annotations for three structures: the left ventricle (LV), myocardium, and right ventricle (RV). The MRI images from both scanners are adopted in our experiments. There is also no patient overlap between the training and testing sets. The train-test ratio is identical to the setting of the MS-CMRSeg Challenge. In addition to the above experiments, four additional experiments are conducted in the ACDC Challenge Dataset, which use 5%, 10%, 20% and 40% of the dataset for training to further validate the proposed approach.

Implementation details. We use original implemented nnU-Net [15] as our backbone with the same augmentations, such as elastic transformation, rotation, scaling, random crop, scaling, adding Gaussian noise, gaussian blur transformation, brightness multiplicative transformation, contrast augmentation transformation, simulate low-resolution transformation, gamma transformation and mirror transformation. Following the suggestion of nnU-Net, the poly learning rate is also introduced in our training. The experiments on MS-CMRSeg Dataset are conducted with a batch size of 12 and a patch size of 512×512. We use a batch size of 44 and patch size of 224 × 320 on the CHAOS Dataset. For the ACDC Challenge Dataset, the batch size and patch size are 56 and 256×224, respectively. The number of both positive and negative pairs are set to 256. We set the fixed high threshold t^h to 0.97, and the temperature τ in the CCDC loss is 0.5. The ratio of $\mathcal{L}_{CE}$, $\mathcal{L}_{DICE}$ and $\mathcal{L}_{CCDC}$ is 1:1:1. Instead of building a memory bank to store the representations for constructing contrastive pairs, representations are sampled dynamically in a batch during each iteration, enabling less memory consumption. This is also demonstrated in [5], where sampling in a mini batch can achieve results similar to those obtained by methods that use additional memory banks.

4.2. Comparison to the State-of-the-Art

Firstly, we benchmark a state-of-the-art *CNN-based model (nnU-Net)* and a *transformer-based method (nnFormer)* in the ACDC Challenge Dataset, the MS-CMRSeg Dataset and the CHAOS Dataset. The results are reported in Tab. 1, Tab. 2 and Tab. 3. nnU-Net serves as our baseline model to validate our CCDC loss since nnU-Net outperforms nnFormer in our experiment settings on three datasets. Compared to the benchmark (nnU-Net), the DSC sees a rise of 1.84%, 3.12% and 3.68% on the three datasets respectively.

Table 1: Segmentation results (DSC and 95HD) on the ACDC Challenge Dataset.

Method	Pubs.	DSC↑			
		RV	myo	LV	mean
nnFormer [27]	2021	78.55	84.31	91.17	84.68
nnU-Net [15]	Nat. Methods 2021	80.90	84.81	91.11	85.61
Focal [31]	ICCV 2017	79.45	84.54	91.12	85.04
GDL [37]	DLMIA 2017	82.10	86.13	91.70	86.64
TopK [38]	NIPS 2017	81.29	85.01	91.34	85.88
MCC [39]	ISBI 2021	80.55	84.54	90.87	85.32
RegionContrast [36]	ICCV 2021	81.38	85.05	91.66	86.03
ContrastiveSeg [6]	ICCV 2021	**83.96**	84.8	91.33	86.70
ReCo [5]	ICLR 2022	82.26	85.87	91.62	86.58
CCDC (ours)	-	83.49	**86.49**	**92.36**	**87.45**
Method	Pubs.	95HD↓			
		RV	myo	LV	mean
nnFormer [27]	2021	6.26	3.43	4.16	4.61
nnU-Net [15]	Nat. Methods 2021	2.87	1.76	3.16	2.60
Focal [31]	ICCV 2017	3.18	1.63	2.61	2.47
GDL [37]	DLMIA 2017	2.92	**1.58**	2.07	2.19
TopK [38]	NIPS 2017	3.16	1.69	2.55	2.47
MCC [39]	ISBI 2021	2.95	2.08	2.89	2.64
RegionContrast [36]	ICCV 2021	2.44	1.87	**1.68**	2.00
ContrastiveSeg [6]	ICCV 2021	2.38	1.91	2.66	2.32
ReCo [5]	ICLR 2022	3.22	1.66	2.3	2.39
CCDC (ours)	-	**2.35**	1.67	1.70	**1.91**
P-values	-	< 5e-2 (DSC), < 5e-2 (HD95)			

4.2.1. ACDC Challenge Dataset

The results for the three classes show that CCDC consistently outperforms other methods of contrastive learning (ReCo +0.87%, ContrastiveSeg +0.75%

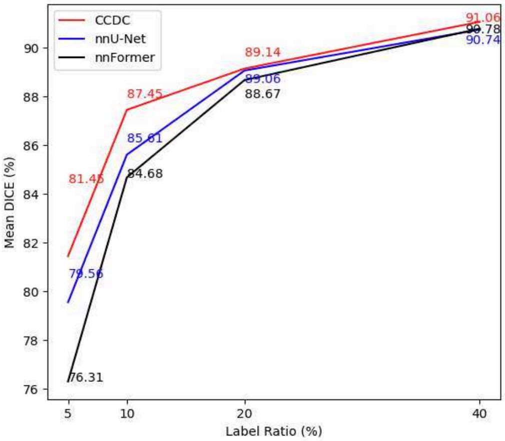

Figure 6: Plot of the mean DSC w.r.t different ratio of the training set over the ACDC Challenge Dataset (5%, 10%, 20%, 40% of the dataset for training), with the blue line representing the result of the baseline: nnU-Net, the black line represents nnFormer, the red line representing our proposed method.

and RegionContrast +1.42%). It may be due to the increased expressiveness of the proposed CCDC model as it relies on optimizing different pixels from easy to hard at different training phases to improve the performance. The average 95HD of all classes is also the smallest, which indicates that performance is improved when CCDC loss is used.

Four additional train-test ratios, 5%, 10%, 20%, 40%, are also evaluated on the ACDC challenge dataset. As shown in Fig. 6, obvious improvements can be observed by comparing the results w/ and w/o CCDC (+1.79%, +1.84%, +0.08%, +0.32%). Notably, our method boosts the baseline model by a significant margin when the amount of the training data is extremely limited. Thus, CCDC demonstrates its potential in facilitating segmentation algorithms in practical applications where labelled data for training is difficult to collect due to concerns about privacy policy and the cost of manual annotations.

4.2.2. MS-CMRSeg Dataset

For MS-CMRSeg dataset, larger patch sizes are adopted, since the height and width of the images in the MS-CMRSeg dataset are much larger compared

Table 2: Segmentation results (DSC and 95HD) on the MS-CMRSeg Dataset.

Method	Pubs.	DSC↑			
		RV	myo	LV	mean
nnFormer[27]	2021	51.96	63.50	70.21	61.89
nnU-Net [15]	Nat. Methods 2021	73.01	74.44	88.95	78.8
Focal [31]	ICCV 2017	74.64	75.26	88.34	79.41
GDL [37]	DLMIA 2017	75.3	73.56	87.57	78.81
TopK [38]	NIPS 2017	72.63	74.22	88.50	78.45
MCC [39]	ISBI 2021	74.08	74.55	87.72	78.78
RegionContrast [36]	ICCV 2021	74.84	76.25	88.51	79.86
ContrastiveSeg [6]	ICCV 2021	75.48	78.44	89.65	81.19
ReCo [5]	ICLR 2022	73.69	75.98	89.23	79.63
CCDC (ours)	-	**76.45**	**79.26**	**90.04**	**81.92**
Method	Pubs.	95HD↓			
		RV	myo	LV	mean
nnFormer[27]	2021	43.89	13.46	13.45	23.60
nnU-Net [15]	Nat. Methods 2021	14.89	9.96	22.79	15.88
Focal [31]	ICCV 2017	15.91	12.6	18.54	15.69
GDL [37]	DLMIA 2017	11.53	9.52	15.12	12.06
TopK [38]	NIPS 2017	13.71	7.57	21.53	14.27
MCC [39]	ISBI 2021	10.69	12.05	24.56	15.76
RegionContrast [36]	ICCV 2021	6.29	5.05	**6.62**	**5.99**
ContrastiveSeg [6]	ICCV 2021	9.54	12.89	24.43	15.62
ReCo [5]	ICLR 2022	6.83	12.18	18.57	12.53
CCDC (ours)	-	**5.60**	**4.31**	13.79	7.90
P-values	-	< 5e-2 (DSC), < 5e-2 (HD95)			

to the ACDC Challenge Dataset (from 256×224 to 512×512). As shown in Tab. 2, CCDC substantially exceeds the state-of-the-art methods by [36, 6, 5], in terms of class-specific DSC scores, our method outperforms the state-of-the-art methods in all classes and the average DSC is significantly improved (2.06% improvement with RegionContrast, 0.73% with ContrastiveSeg and 2.29% with ReCo on DSC). This demonstrates that CCDC is consistently beneficial for the baseline model, whether the training data are small images or large images with rich details. We also show class-specific results in the 95HD metrics. The 95HD is the smallest in the class of RV and myo and the average 95HD of CCDC also ranks second best.

4.2.3. CHAOS Dataset

We evaluated our method using the CHAOS dataset (four classes segmentation tasks). Tab. 3 shows the results of the segmentation for the liver, RK, LK and spleen. We can see that all contrastive learning methods [36, 6, 5] show improvements over baseline on DSC, demonstrating the benefits of contrastive learning. Our CCDC learning with a dynamic learning strategy consistently keeps improving on the average DSC score. These results demonstrate the effectiveness of our dynamic learning strategy with extremely limited data (6 images from 2 patients only) and validate that our dynamic sampling strategy can be applied to a wider range of multi-class segmentation tasks. The 95HD of the CCDC(ours) algorithm is not optimal, and the DSC of the liver and spleen are not optimal. We have conducted further analysis and found that it may be due to the dataset we used in the experiment contains more classes and many images that contain a high degree of variability in terms of size, shape, and texture of the structures of interest. They together make it challenging to achieve optimal results for all classes in multi-class tasks, where the 95HD metric might not be as relevant as the DSC. This may be a result of the class imbalance or overlapping features among classes. Despite these challenges, our overall performance on the DSC metric has still improved.

Table 3: Segmentation results (DSC and 95HD) on the CHAOS Dataset.

Method	Pubs.	DSC↑				
		liver	RK	LK	spleen	mean
nnFormer[27]	2021	70.88	58.41	56.36	62.66	62.08
nnU-Net [15]	Nat. Methods 2021	73.79	57.09	63.27	60.27	63.61
Focal [31]	ICCV 2017	76.31	59.28	67.43	61.95	66.24
GDL [37]	DLMIA 2017	74.8	54.65	65.17	61.26	63.97
TopK [38]	NIPS 2017	75.85	55.14	68.63	**62.68**	65.57
MCC [39]	ISBI 2021	75.63	56.08	63.88	61.09	64.17
RegionContrast [36]	ICCV 2021	74.45	57.15	68.58	57.40	64.39
ContrastiveSeg [6]	ICCV 2021	**77.52**	56.59	66.11	57.36	64.40
ReCo [5]	ICLR 2022	75.44	61.85	67.66	61.05	66.50
CCDC (ours)	-	75.03	**62.43**	**70.81**	60.89	**67.29**
Method	Pubs.	95HD↓				
		liver	RK	LK	spleen	mean
nnFormer[27]	2021	27.58	29.24	30.35	28.99	29.04
nnU-Net [15]	Nat. Methods 2021	17.80	37.93	30.16	21.19	26.77
Focal [31]	ICCV 2017	13.86	23.05	19.6	16.85	18.34
GDL [37]	DLMIA 2017	14.16	20.30	30.68	25.86	22.75
TopK [38]	NIPS 2017	14.49	14.62	9.90	36.77	18.95
MCC [39]	ISBI 2021	13.27	20.31	25.28	**16.04**	18.72
RegionContrast [36]	ICCV 2021	14.75	**12.74**	**7.34**	25.01	**14.96**
ContrastiveSeg [6]	ICCV 2021	**12.45**	19.21	22.03	33.80	21.87
ReCo [5]	ICLR 2022	16.31	20.90	24.94	23.17	21.33
CCDC (ours)	-	14.51	20.23	23.97	19.64	19.59
P-values	-	$<$ 5e-2 (DSC), $<$ 5e-2 (HD95)				

Table 4: Dynamic query pixels sampling strategy on ACDC Challenge Dataset.

Method	Dynamic Query Sampling		DSC↑			
	Adaptive T	Adjusted Ratio	RV	myo	LV	mean
W/O Contrastive Learning			80.90	84.81	91.11	85.61
Pixel-level	-	-	82.13	86.15	91.81	86.70
	√	-	82.39	85.92	91.78	86.70
	-	√	82.60	85.80	92.15	86.85
	√	√	83.84	85.82	91.64	**87.10**
Region-level	-	-	80.85	85.76	91.72	86.11
	√	-	82.21	86.86	92.38	87.15
	-	√	82.68	85.59	91.84	86.71
	√	√	83.49	86.49	92.36	**87.45**

4.3. Ablation Studies

4.3.1. The Impact of Dynamic Query Sampling Strategy with Pixel-Level and Region-Level Representations

Tab. 4 shows the ablation study with and without the dynamic contrastive learning in easy and hard subsets which are separated by an adaptive threshold (adaptive T) with the adjusted ratio for sampling easy to hard queries on the ACDC challenge dataset.

We examined our proposed dynamic contrastive learning on pixel-level representations and region-level representations separately. It is worth noting that despite not using dynamic sampling strategies, contrastive learning improves performance with and without the representation enhancement strategy. More notably, when using a fixed threshold to build easy and hard sets, the introduction of dynamic contrastive learning to adjust easy-hard ratios improves the results by using the pixel-level or region-level representations (Pixel-level: 86.7% to 86.85%, Region-level: 86.11% to 86.71%). Fixing the ratios of sampling easy and hard queries and using the changed threshold to select easy and hard sets to obtain a similar average score (86.7%) in pixel-level contrastive learning, but

1.04% (86.11% to 87.15%) improvement at the region level. This shows that our the dynamic sampling strategy can achieve some performance improvement regardless of adjusting the threshold of the easy and hard sets or their sampling ratio. When using the adaptive threshold and the adjusted sampling ratio together to separate easy and hard sets, the performance is further improved, suggesting that the dynamic approach can capture the pixels that are most in need of optimisation at different training phases to improve segmentation performance.

Table 5: Query sampling from different intra-class subsets on ACDC Challenge Dataset.

Method	Sampling Strategy	DSC↑			
		RV	myo	LV	mean
W/O Contrastive Learning		80.90	84.81	91.11	85.61
Pixel-level	Hard	80.29	85.77	90.86	85.64
	Easy	81.87	85.66	92.17	86.57
	Easy to Hard	81.67	85.27	91.47	86.14
	Dynamic Sampling	83.84	85.82	91.64	**87.10**
Region-level	Hard	82.34	86.14	92.01	86.83
	Easy	82.22	86.95	92.38	87.18
	Easy to Hard	83.87	86.12	91.86	87.28
	Dynamic Sampling	83.49	86.49	92.36	**87.45**

4.3.2. The Impact of Query Sampling from Different Subsets

Here, we focus on different strategies for sampling query pixels: 1) sampling queries in the easy set; 2) sampling queries in the hard set; 3) sampling queries from easy to hard; 4) dynamically changing strategies for queries sampling. As described in the methodology section, we do not use fixed thresholds to identify easy and hard sets, but instead use an adaptive threshold to separate the easy and hard queries to be sampled based on how well the network is trained during different phases. The pixel-level representations and the region-level representations are both used for the sampling strategy comparison. Tab. 5 shows the results of the four sampling strategies on the ACDC challenge dataset. The four aforementioned sampling strategies yield better results than the benchmark. The results of the easy sampling strategy consistently outperform those of the hard-only sampling (Pixel-level: 85.64% to 86.57%, Region-level: 86.83% to 87.18%). The proposed dynamic sampling strategy further improves the performance (Pixel-level: +0.53%, Region-level: +0.27%). It verified that using only hard pixels for training may result in the network not converging consistently. Training with only easy pixels can achieve better performance than only hard pixels, but the performance is also lower than dynamic sampling strategy. This is because using only easy pixels does not provide enough gradient descent for optimization, which may influence performance improvement. It can be seen that the performance of the easy-to-hard sampling strategy with pixel-level representations decreased when compared with only using easy pixels, this may be because easy-to-hard sampling does not measure the current performance of the model and may not necessarily provide the pixels needed by the current model. However, the performance of the easy-to-hard strategy with region-level representations is better than using only easy or hard pixels, but not as good as our dynamic sampling strategy. As our dynamic sampling strategy can dynamically change the sampling strategy based on the performance of the model and can select the most representative pixels at different training stages. The result reveals that our method can, not only improve the basic generalization ability when the performance is low, but also enhance the hard problem-solving ability when the performance is strong, suggesting that the approach is beneficial for seeking better pixel-level segmentation applications.

4.3.3. The Impact of Hard Negative Class Mining Strategy

Tab. 6 shows the performance improvement of the ACDC Challenge Dataset for different negative class sampling strategies. Despite using different sampling strategies, we find that CCDC loss has a consistent improvement of 1%-2%. Uniform sampling results in 0.29% improvement compared to random sampling. This may be due to the class imbalance problem, if a class in a mini-batch has

Table 6: Hard negative class mining strategy on ACDC Challenge Dataset

Method	DSC↑			
	RV	myo	LV	mean
Baseline	80.9	84.81	91.11	85.61
Random Sampling	81.91	86.29	92.15	86.78
Uniform Sampling	84.12	84.78	92.30	87.07
Dynamic Sampling	83.49	86.49	92.36	**87.45**

a small number of pixels, and this class is also hard to distinguish, random sampling is not sufficient when sampling from this hard class for training, resulting in missing samples and low performance on this hard class. This can also be demonstrated by the results of the hard but fewer pixels RV class (DSC of Random: 81.91% and DSC of UniForm 84.12%). Following our class confusing degree-guided hard-negative class mining strategy, negative samples can be sampled from each class with different ratios, which is more beneficial in sampling negative samples thus obtaining significant performance improvement.

4.3.4. The Impact of Region Enhancement Strategy

Table 7: Region enhancement strategy on the ACDC Challenge Dataset.

Number of Pixels	DSC↑			
	RV	myo	LV	mean
1	83.84	85.82	91.64	87.10
2	83.73	86.30	91.36	87.13
4 (Stride=2)	83.05	86.36	92.60	87.34
4 (Stride=1)	83.49	86.49	92.36	**87.45**

We evaluated the effectiveness of the region enhancement strategy. Different region enhancement schemes are adopted and shown in Fig. 5. Specifically, using different numbers of pixels and strides in diagonal, horizontal and vertical directions to build regions can achieve different improvements. As shown in Tab. 7, the region enhancement strategy can achieve a consistent improvement over the benchmark, and it is observed that the performance is improved with the increase of the number of neighbouring pixels. The closer the neighboring pixels are to the center pixel, the further the performance can be improved. This is because pixels in boundary regions are often more difficult to distinguish and require stronger local location information for fine differentiation. In contrast, our representation enhancement strategy introduces local location information from the neighbouring pixels for contrastive learning, thus providing a supplement to determine the cross-class boundaries.

4.3.5. The Impact of the Fixed High Threshold to Choose Confidence Core

Table 8: Different fixed high threshold on ACDC Challenge Dataset.

High Threshold	DSC↑			
	RV	myo	LV	mean
W/O Dynamic	80.90	84.81	91.11	85.61
0.50	83.27	86.00	92.39	87.22
0.75	82.78	85.50	92.46	86.91
0.90	82.97	86.50	92.53	87.33
0.97	83.49	86.49	92.36	**87.45**

We evaluated the impact of a fixed high threshold used to select the confidence core. In Tab. 8, it can be observed that slightly better performance can be obtained with a relatively large threshold (0.97 and 0.9). When the confidence threshold drops to 0.75 and 0.5, compared with using the thresholds over 0.9, performance declined to a 87.22% and 86.91%. However, compared to the baseline, the performance is still greatly improved. We can also observe that our method is not highly dependent on fixed high-threshold hyperparameters for confidence pixel selection. This is because a fixed threshold is needed to ensure that the selected pixel can be predicted as the correct class. Thus, the pixels above this threshold can form a class-specific prototype to guide the learning of other low-confidence pixels within the class.

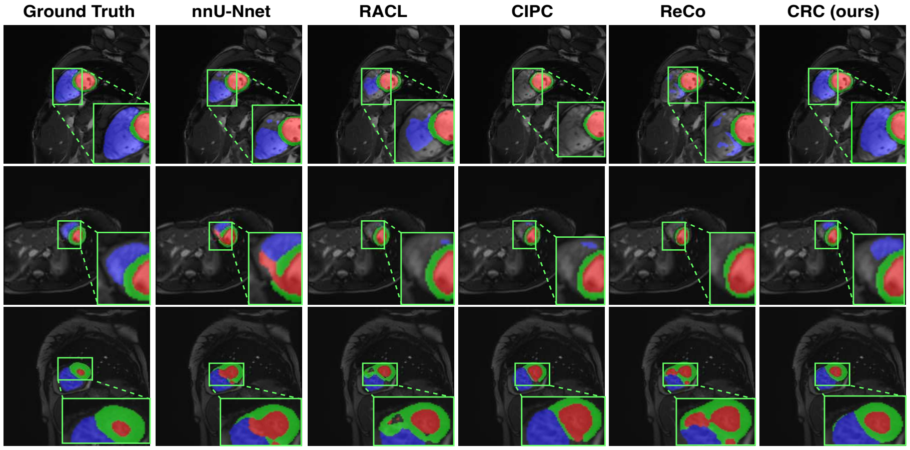

Figure 7: Typical segmentation result comparing different approaches on the ACDC Challenge Dataset. Blue, green and red denote the classes of RV, myocardium, and LV, respectively.

5. Visualization

Fig. 7 shows the segmentation results using a model trained with 10% data on the ACDA dataset when compared with the state-of-the-art contrastive learning methods [36, 6, 5]. The prediction results of the RV class in the first row and the myocardium class in the second and third rows show that our method achieves significant improvements in the identification of boundaries and the segmentation of hard classes. This is attributed to the dynamic learning strategy which identifies the core representations of a class and pulls different hardness pixels to the core along with the training, Furthermore, negative class mining also identifies the hard negative classes, avoiding a large number of pixels being misclassified to the hard negative classes. The representation enhancement strategy introduces local spatial information, further refining the determining of boundaries and reducing the confusion between different adjacent classes.

6. Conclusion and Discussion

In this paper, we present a dynamic contrastive learning framework for medical image segmentation that leverages class-confidence and confusion. The framework features a novel loss function, the CCDC loss, which dynamically samples contrastive pixels based on the model's performance. Our approach selects the most expressive pixels as positive and negative pairs during different training phases and employs a hard negative class mining strategy to enhance effectiveness. The results show that our method outperforms the state-of-the-art on three challenging datasets and has potential for other multi-class tasks such as classification and detection. While our approach has achieved significant progress, we acknowledge the need to address the dynamic adjustment of easy/hard boundaries in situations where the dataset is challenging. To enhance robustness, we suggest potential future research directions, including implementing a stopping criterion. Future work also includes exploring the method's applicability to other domains or tasks, scalability, efficiency, robustness, and generalization. Another avenue for future research is to investigate combining our method with other techniques to further improve performance.

References

[1] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[2] Z. Cao, W. Wang, L. Huo, S. Niu, Unsupervised class-to-class translation for domain variations, Pattern Recognition 138 (2023) 109346.

[3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[4] A. Ben Saad, K. Prokopetc, J. Kherroubi, A. Davy, A. Courtois, G. Facciolo, Improving pixel-level contrastive learning by leveraging exogenous depth information, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2380–2389.

[5] S. Liu, S. Zhi, E. Johns, A. J. Davison, Bootstrapping semantic segmentation with regional contrast, in: International Conference on Learning Representations, 2022.

[6] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7303–7313.

[7] T. Sun, C. Lu, T. Zhang, H. Ling, Safe self-refinement for transformer-based domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7191–7200.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE transactions on pattern analysis and machine intelligence 32 (9) (2010) 1627–1645.

[9] G. Hacohen, D. Weinshall, On the power of curriculum learning in training deep networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 2535–2544.

[10] A. Graves, M. G. Bellemare, J. Menick, R. Munos, K. Kavukcuoglu, Automated curriculum learning for neural networks, in: international conference on machine learning, PMLR, 2017, pp. 1311–1320.

[11] J. L. Elman, Learning and development in neural networks: The importance of starting small, Cognition 48 (1) (1993) 71–99.

[12] Y. Wang, W. Gan, J. Yang, W. Wu, J. Yan, Dynamic curriculum learning for imbalanced data classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5017–5026.

[13] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769.

[14] X. Zhang, Q. Guo, Y. Sun, H. Liu, G. Wang, Q. Su, C. Zhang, Patch-based fuzzy clustering for image segmentation, Soft Computing 23 (9) (2019) 3081–3093.

[15] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2) (2021) 203–211.

[16] X. Zhuang, J. Xu, X. Luo, C. Chen, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar, et al., Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge, Medical Image Analysis 81 (2022) 102528.

[17] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?, IEEE transactions on medical imaging 37 (11) (2018) 2514–2525.

[18] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al., Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation, Medical Image Analysis 69 (2021) 101950.

[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[20] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[21] K. Wang, X. Zhang, Y. Lu, W. Zhang, S. Huang, D. Yang, Gsal: Geometric structure adversarial learning for robust medical image segmentation, Pattern Recognition 140 (2023) 109596.

[22] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, D. Yang, Eanet: Iterative edge attention network for medical image segmentation, Pattern Recognition 127 (2022) 108636.

[23] J. Chen, J. Zhang, K. Debattista, J. Han, Semi-supervised unpaired medical image segmentation through task-affinity consistency, IEEE Transactions on Medical Imaging.

[24] J. Chen, H. Li, J. Zhang, B. Menze, Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac mr images segmentation, in: Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10, Springer, 2020, pp. 317–325.

[25] J. Chen, W. Li, H. Li, J. Zhang, Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer, 2020, pp. 187–196.

[26] Q. Yan, S. Liu, S. Xu, C. Dong, Z. Li, J. Q. Shi, Y. Zhang, D. Dai, 3d medical image segmentation using parallel transformers, Pattern Recognition 138 (2023) 109432.

[27] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: Interleaved transformer for volumetric segmentation, arXiv preprint arXiv:2109.03201.

[28] Z. Liu, F. Wu, Y. Wang, M. Yang, X. Pan, Fedcl: Federated contrastive learning for multi-center medical image classification, Pattern Recognition (2023) 109739.

[29] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard negative mixing for contrastive learning, Advances in Neural Information Processing Systems 33 (2020) 21798–21809.

[30] K. Yuan, G. Schaefer, Y.-K. Lai, Y. Wang, X. Liu, L. Guan, H. Fang, A multi-strategy contrastive learning framework for weakly supervised semantic segmentation, Pattern Recognition 137 (2023) 109298.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[32] M. S. Hossain, J. M. Betts, A. P. Paplinski, Dual focal loss to address class imbalance in semantic segmentation, Neurocomputing 462 (2021) 69–87.

[33] X. Wang, Y. Chen, W. Zhu, A survey on curriculum learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9) (2021) 4555–4576.

[34] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. A. González Ballester, G. Piella, Medical-based deep curriculum learning for improved fracture classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 694–702.

[35] X. Yao, X. Feng, J. Han, G. Cheng, L. Guo, Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning, IEEE Transactions on Geoscience and Remote Sensing 59 (1) (2020) 675–685.

[36] H. Hu, J. Cui, L. Wang, Region-aware contrastive learning for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16291–16301.

[37] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[38] Y. Fan, S. Lyu, Y. Ying, B. Hu, Learning with average top-k loss, Advances in neural information processing systems 30.

[39] K. Abhishek, G. Hamarneh, Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 225–229.

Dynamic contrastive learning guided by class confidence and confusion degree for medical image segmentation

Jingkun Chen[a,b], Changrui Chen[b], Wenjian Huang[a], Jianguo Zhang[a,d,*], Kurt Debattista[b,*], Jungong Han[b,c,*]

[a] *Department of computer science and engineering, Southern University of Science and Technology, Shenzhen, 518055, China*
[b] *WMG Visualization, University of Warwick, Coventry, CV4 7AL, United Kingdom*
[c] *Department of Computer Science, University of Sheffield, Sheffield, S10 2TN, United Kingdom*
[d] *Peng Cheng Lab, Shenzhen, China, 518000, China*

Abstract

This work proposes an intra-Class-confidence and inter-Class-confusion guided Dynamic Contrastive (CCDC) learning framework for medical image segmentation. A core contribution is to dynamically select the most *expressive pixels* to build positive and negative pairs for contrastive learning at different training phases. For the positive pairs, dynamically adaptive sampling strategies are introduced for sampling different sets of pixels based on their hardness (namely the easiest, easy, and hard pixels). For the negative pairs, to efficiently learn from the classes with high confusion degree w.r.t a query class (i.e., a class containing the query pixels), a new *hard class* mining strategy is presented. Furthermore, pixel-level representations are extended to the neighbourhood region to leverage the spatial consistency of adjacent pixels. Extensive experiments on the three public datasets demonstrate that the proposed method significantly surpasses the state-of-the-art. The code is publicly available at https://github.com/jingkunchen/ccdc.

Keywords: `Class Confusion Degree, Dynamic Contrastive Learning, Medical Image Segmentation`

[⋆] This work is supported in part by National Key Research and Development Program of China (2021YFF1200800) and National Natural Science Foundation of China (Grant No. 62276121)

[*] Corresponding authors

1. Introduction

Deep learning based medical image segmentation algorithms are capable of providing clinical guidance under the right conditions, such as surgery planning and post-surgery monitoring. However, training a successful deep-learning-based segmentator requires a large amount of pixel-level annotation data, which is always rare and expensive, especially for 3D images. It remains a challenge to train a satisfactory deep learning-based segmentation algorithm with limited annotated data.

In recent years, contrastive self-supervised representation learning is widely adopted to reduce the use of labelled data, which turns out to be successful in solving *image-level* tasks, such as classification [1] and image translation [2], and *instance-level* tasks, such as detection [3]. In contrastive learning, constructions of positive (similar) pairs and negative (dissimilar) pairs [4] are essential. This is not trivial, especially when dealing with a large number of pixels, e.g., segmentation task, as naive contrastive learning that exhaustedly combines all pixels becomes computationally prohibitive. Therefore, sampling a portion of pixels for contrastive learning in segmentation task is usually adopted as an alternative [5].

To build positive pairs in contrastive learning, strategies of sampling pixels from images for segmentation task, such as random sampling [6] and active hard sampling [5], suffer from several limitations. More concretely, random sampling [6] treats each pair equally without considering the effectivenesses of the selected pixels. Often, a large pixel sampling ratio will inevitably introduce redundant information, as not all pairs are needed at the same time. For example, easy pixels cannot contribute to model optimisation when the performance of the model is high. Alternatively, active hard sampling strategies [5] focus only on hard pixels but exclude easily distinguished samples during all training phases. This may lead to hindered convergence due to the sharp gradient fluctuations when feeding hard pixels to a model with unstable/or poor performance [7]. Compared to using hard pixels, the *non-hard* pixels are more conducive to training when the performance of a model is poor.

On the other hand, for negatives sampling, pixels sampled from an easy class tend to contribute less to the optimisation of a model, especially in the case of class imbalance (easily distinguishable classes contain more pixels, while hard classes contain fewer pixels) [8]. The sampled negative pixels are more useful for learning discriminative feature only when the pixels from hard classes are sampled frequently.

Thus, two to-be-solved problems appear when applying pixel-level contrastive learning in segmentation task: Firstly, as the number of pixels in a dataset is huge, efficiently sampling representative pixels is the main problem to be solved. Secondly, the needs of pixels of a model will change at different phases. It is more reasonable to use appropriate sampling strategies during different training phases. Solving these questions can provide insights to promote the design of an efficient pixel-level contrastive learning sampling strategy. We first investigate the first question by exploring the existing pixel-level contrastive learning methods, which sample pixels from a class to construct positive pairs (z and z^+) and different classes to construct negative pairs (z and z^-). However, most existing efforts treat all pixels equally and do not consider the hardness of the pixels. In an image containing regions from different classes, the center pixels in one class-specific region tend to be easily distinguishable than the border pixels, as border pixels could be easily confused with neighbouring pixels belonging to other classes. Thus, to separate different regions within a class and treat the pixels in these regions differently, it is necessary to divide the pixels in a class into different groups according to their hardness and each group should be treated differently. For the second question, most of the works in the literature use the same sampling strategy at the different training phases. However, if there are significant changes in the performance of the model, a fixed sampling strategy cannot always sample representative pixels to contribute to the model. For example, when the performance of the model is already good enough, extracting a large number of easy pixels does not contribute much to the optimisation. On

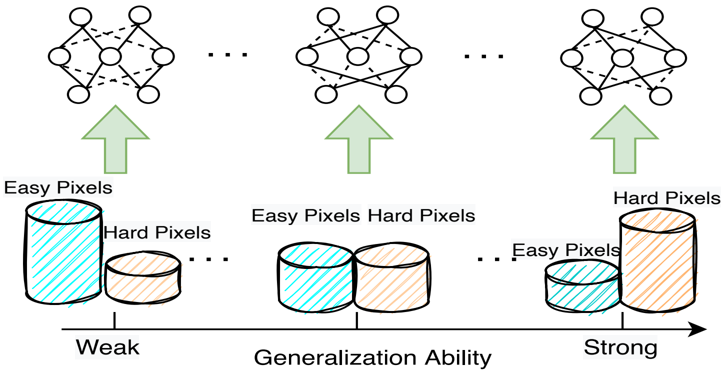

Figure 1: Our intra-class confidence degree guided dynamic contrastive learning aims to select different sampling strategies to build positive pairs at different training phases. The dashed lines with arrows denote the data flow; the solid lines are used to indicate the construction of the positive pairs. The height of the cylinder represents the number of easiest, easy and hard pixels in Class A in the current training phase. The length of the rectangle indicates the number of samples. When the intra-class confidence degree of a model is low, a small sampling ratio to the hard pixels and a much larger sampling ratio to the easy pixels is applied for training. When the intra-class confidence degree of a model is strong, the sampling strategy gradually moves to the opposite ratio for them.

the contrary, taking more hard pixels when the performance of a model is low will produce a large gradient, which may prevent the model from converging efficiently.

Based on the above discussion, we first propose to dynamically change the sampling strategy at different training stages to select the most expressive pixels. Based on this idea, a *dynamic intra-class confidence degree* is introduced to divide the pixels into three groups: query pixels (z), intra-class core (z^+) and inter-class negative pixels (z^-). Query pixels (z) and the intra-class core (z^+) are used to construct positive pairs, and query pixels (z) and inter-class negative pixels (z^-) are used to construct negative pairs. In the following part of this paper, we use the pixels of a class A as an example to illustrate our contrastive strategy. For sampling queries and core, as shown in Fig. 1, several adaptive sampling strategies based on variations in model performance (low, medium, and high confidence degree for Class A) are introduced to select pixels. Specifically,

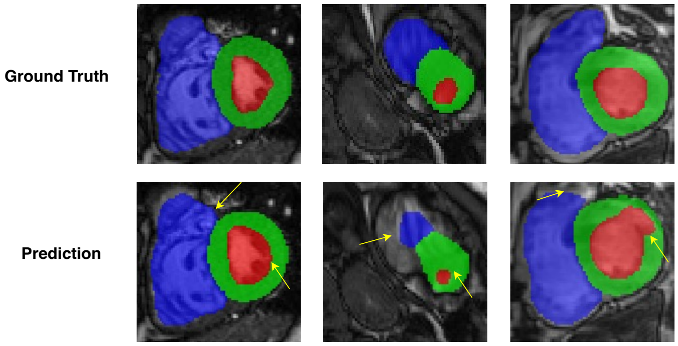

Figure 2: Cardiac images with ground truth masks, the left and right ventricles (Blue and Red regions) are two organs that are separated by the myocardium (Green region) and not adjacent to each other. Pixels in the ventricles class are often misclassified as myocardium class because pixels on the boundary are easily affected by other classes.

for all phases, the *easiest* distinguishable pixels of class A can be considered as the core (z^+). The representations generated by the core (z^+) can be used as a prototype to boost other query pixels (z) in the same class for class-specific representation learning [9]. When the overall classification accuracy of the pixels in class A is low, the *easier* pixels in class A should be sampled more as queries for optimisation to reduce learning difficulty and improve overall performance [10]. When the accuracy of the pixels in class A is medium, the sampling ratio of the *easier* query pixels should be reduced and the *harder* ones should be further considered, i.e. starting easy and gradually presenting more complex concepts [11] can obtain better learning effects. If accuracy is high, more attention should be paid to the *hard* pixels to improve the hardest problem-solving skills. It is worth noting that our approach is different from the easy-to-hard sampling strategy [12], as we dynamically measure the performance of the model on each class at different phases and adapt accordingly. The proposed method can dynamically change the sampling strategy to sample the most representative pixels to train a more robust model (verified in Sect. 4.3.2) and avoid the risk of the model crashing due to a catastrophic incident (the great variance of gradients) by the sampled hard pixels [7].

Generally, pixels sampled from more confusing classes as negatives can improve the efficiency of network optimisation, as hard classes pixels allow the model to focus on distinguishable details [13]. Since pixels from two adjacent classes or two similar classes are more difficult to distinguish, and pixels from two distant classes or dissimilar classes are often easier to distinguish, it is possible to focus on relatively difficult classes to get better performance. For example, as shown in Fig. 2, the right and left ventricles are two organs that are separated by the myocardium and not adjacent to each other. For the right and left ventricles, because of their different structures, colour and intensity distribution, the right ventricle can be easily separated from the left ventricle. But for the myocardium class, the neighbouring pixels in the myocardium class greatly influence the ventricle classes in adjacent regions. Based on this observation, negative pixel sampling strategies should be designed *differently* from query sampling strategies which need to choose easy pixels when the performance is low and hard pixels when the performance is high. The proposed method helps determine which class is more confusing. The *inter-class-confusion degree* is introduced to measure the *inter-class* difficulty from one class to other classes at different phases. The inter-class difficulty can be used to control the number of sampled pixels from different negative classes.

Finally, in the segmentation task, local relationship is crutial for determining the boundary of two classes. A representation at the positive is definitely influenced by its neighbouring pixels [14]. We enhance the pixel-level representation by combining the CNN features of individual pixels with their neighbouring pixels. Experiments show that our region-based contrastive learning method is more effective than using the single pixel-level one.

In this paper, a new loss, intra-Class-confidence and inter-Class-confusion guided Dynamic Contrastive (CCDC) loss, is introduced to dynamically change the pixels sampling strategies at different training phases. We use the state-of-the-art nnU-Net [15] framework as the backbone to evaluate our approach on three datasets, the MS-CMRSeg Dataset [16], the ACDC Challenge Dataset [17] and CHAOS (combined (CT-MR) Healthy Abdominal Organ Segmentation)

Dataset [18]. The proposed CCDC learning outperforms the state-of-the-art methods. The contributions of this paper are summarized below.

- We design a *intra-class-confidence and inter-class-confusion* guided dynamic contrastive framework and introduce a new loss, CCDC loss, for medical image segmentation. This is the first attempt to dynamically change the sampling strategy to select the most *expressive pixels* to construct positive and negative pairs based on the current circumstance of each class.

- A *intra-class-confidence degree* is introduced to measure the performance of a model on each class. A dynamically adaptive intra-class sampling strategy based on the intra-class-confidence-degree is used to separate the easiest, easy, and hard pixels for intra-class similarity learning.

- To learn the class differences, a *inter-class-confusion degree* is proposed to dynamically adjust the sampling ratio of the positive and negative classes, and prioritize sample pixels from the hard classes.

- We extend pixel-level contrastive learning to the region-level by considering the spatial dependency of neighboring pixels, which turns out to be more suitable for the segmentation task.

- The results of three public datasets show that the proposed method is more effective than other pixel-level contrastive learning methods, which is highly relevant in the field of semantic segmentation, especially in medical imaging where accurate segmentation is crucial for diagnosis and treatment.

2. Related Work

Although traditional computer vision techniques have proven to be effective in extracting representations from image data, the rise of Convolutional Neural

Networks (CNNs) has challenged their dominance and established a new benchmark for representation learning. The ability of CNNs to learn hierarchical representations from raw data, incorporating both local and global context, has been demonstrated through numerous studies and outperforms traditional computer vision methods in many tasks. In recent years, CNN-based approaches such as FCN [19], U-Net [20] and other related models [21] have gained significant popularity for medical image segmentation across various fields, including tasks such as lung nodules segmentation [22], hippocampus segmentation [23], ventricle and myocardium segmentation [24] and MRI-CT whole heart segmentation [25]. Furthermore, the application of Transformers in medical image segmentation has been on the rise as well, with notable models like EffTrans[26] and nnFormer [27]. For CNN- and transformer-based backbones, most segmentation models use pixel-level cross-entropy loss and Sorensen-Dice coefficient loss to train the networks. However, none of these loss functions takes explicitly into account the relationship between different pixels (whether the pixels come from the same classes), which can be fundamental for a segmentation task.

2.1. Pixel-level Contrastive learning

Contrastive learning is a form of comparative learning that involves forming positive pairs and negative pairs to learn the common representation between similar instances and discriminative representations between non-similar instances [28]. For classification problems, contrastive learning is generally performed on image-level representations. Positive samples are sampled from homologous images to ensure similarity, while negative samples are usually derived from non-homologous images. Generally, negative mining strategies in contrastive learning are widely used for image-level classification tasks to reduce computational effort and improve efficiency [29].

In the segmentation task, to exploit the potential of the labelled images, contrastive learning helps capture detailed information to consider the elaborate pixel-level representations rather than the whole image-level representation [30]. Although there have been some works on pixel-level contrastive learning, these methods have the following limitations:

(1) Sampling strategies such as *random sampling, uniform sampling* , are usually unstable and may result in missing necessary samples or taking unnecessary samples. *Hard pixel mining* strategies [6] focus on the hardest pixels and ignore the contribution of non-hard pixels. When the performance of a model is low, *non-hard* pixels should be given priority over the hardest ones to avoid the risk of model divergence due to the great variance of gradients by a large loss [7]. *Balancing weight of the easy and hard pixels*, such as Focal loss [31], gives less weight to easy pixels and gives more weight to hard misclassified pixels. However, It tends to produce vanishing gradients during backpropagation, penalizes negative classes in reverse, or employ non-optimal loss weights between classes [32].

(2) In the segmentation task, recent works ignore the spatial dependency between neighbouring pixels. A pixel-level representation has a strong spatial dependency on its neighbouring pixels. The pixels' spatial dependency in segmentation tasks is crucial to delineating the boundary. Treating all pixels individually would lose the spatial information between the pixels [14].

2.2. Dynamic Learning

For dynamic learning, the traditional approach is to learn from easy to hard, designed to mimic the meaningful learning strategies of human learning [33]. The basic idea of easy to hard learning is to initially train the model with simple data and then gradually feed more difficult samples to obtain a better generalization [12]. This strategy has shown potential for image-level classification [34] and object-level detection [35]. However, there are few studies on easy-to-hard strategies for pixel-level contrastive learning. Meanwhile, directly using the easy-to-hard learning strategy for pixel-level contrastive learning still suffers from limitations due to pixels sampling risk; a model with low performance caused by insufficient training with easy pixels may make model more unstable or even cause model to crash when hard pixel oversampling. Since learning with hard samples in a weak model introduces a larger uncertainty [7], sampling from easy to hard can reduce the risk of model collapse and build a robust model. However, continuing to sample harder pixels may be counterproductive when the model's performance decreases. Compared to the easy-to-hard strategy, as shown in Sect. 4.3.2, our adjusting sampling strategy dynamically samples different difficulty pixels based on the current performance of the model and can perform well in representation learning.

3. Methodology

The quality of the pixel-level representations is crucial to the performance of pixel-level segmentation. In this study, to improve the quality of the pixel-level representations, a dynamically pixel-level contrastive learning strategy, termed intra-class-confidence and inter-class-confusion guided dynamic contrastive learning, is designed to learn pixel-level representations with stronger intra-class similarity and inter-class distinguishability.

As shown in Fig. 3, we employ nnU-Net [15] as the backbone, the network is responsible for extracting feature maps from the input image, and the feature maps are denoted by h_i in the figure, followed by two heads: (1) one segmentation head with a Softmax function, the segmentation head is used for predicting the class label of each pixel and (2) one projection head is then applied on h_i to project them into a feature space where contrastive learning is performed, the feature maps after the projection head are denoted by z_i in the figure. The cross-entropy loss ($\mathcal{L}_{CE}$) and DICE loss ($\mathcal{L}_{DICE}$) are applied to the segmentation head branch. In the projection head, we introduce our proposed dynamic learning strategy for contrastive learning.

3.1. Pixel Grouping In Contrastive Learning

To conduct pixel-level contrastive learning, we first divide all pixels of images in a batch into three groups: query pixels (z), and intra-class core (z^+, which is used to build a class-specific prototype) and inter-class negative pixels (z^-, which are used for learning the class-distinguished details). In our contrastive learning framework, query pixels (z), and intra-class core (z^+) are used

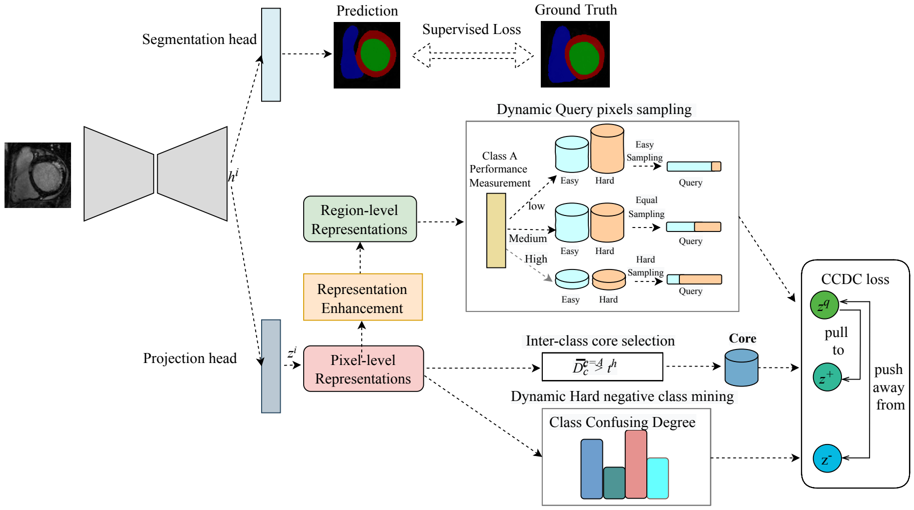
Segmentation head
Prediction
Supervised Loss
Ground Truth
h^i
Dynamic Query pixels sampling
Class A Performance Measurement
low
Medium
High
Easy
Hard
Easy Sampling
Equal Sampling
Hard Sampling
Query
Region-level Representations
Representation Enhancement
Projection head
z^i
Pixel-level Representations
Inter-class core selection
$\overline{D}_c^{c=A} > t^h$
Core
Dynamic Hard negative class mining
Class Confusing Degree
CCDC loss
z^q
pull to
z^+
push away from
z^-

Figure 3: Model overview. Two modules are designed to enhance the feature learning of the backbone of the auto-encoder. One is for supervised learning against the ground truth and the other is for intra-class confidence degree and inter-class-confusion degree guided dynamic contrastive learning, with a projection head mapping the feature representations to the space where contrastive loss is applied.

to construct positive pairs, and query pixels (z) and inter-class negative pixels (z^-) are used to construct negative pairs.

3.2. Confidence Class-Specific Core

To sample pixels to build positive pairs for contrastive learning, we start by defining the *intra-class confidence degree*, which will be used to measure the performance of a model at different training phases. We first use the class-specific ground-truth mask to split all the pixels into different classes according to their ground truth class. In the following part of this section, we use the pixels of the class A as an example to illustrate our dynamic contrastive sampling strategy, where A is the ground truth class. For i-th pixel in a class, we denote the predicted probability of this class as $p_i^c = P(y_i = c|x_i)$, where c is the predicted class, and, y_i and x_i are the predicted label and pixel i in image x respectively. When $c = A$, we define a *intra-class confidence degree* by aggregating the probabilities $p_i^{c=A}$ for all pixels with respect to class A:

$$\overline{D}_{A}^{c=A} = avg(\{p_i^{c=A} | i \in \mathcal{I_A}\}), \tag{1}$$

where $\mathcal{I}_A$ denotes the set of indices of all pixels in class A and $p_i^{c=A}$ is the probability pixel i predicted to belong to class A and avg is the average operation.

As shown in Fig. 1, we select the most representative pixels in class A as the core by intra-class confidence degree above a high threshold, as these pixels are the easiest to distinguish. The core is designed to be a class-specific prototype of the class A which can not only be used to distinguish this class from other classes but also can boost other intra-class pixels' learning. The core for the class A is defined as the *fused representation* produced by averaging these high confidence pixels' representations with predicted probabilities being larger than a threshold: $p_i^{c=A} > t^h$. Let z_i be the representation of pixel i. The core z^{A+} of the class A is calculated as

$$z^{A+} = \frac{\sum_i^{\mathcal{I_A}} \alpha_i z_i}{\sum_i^{\mathcal{I_A}} \alpha_i}, where\ \alpha_i = \mathbb{1}(p_i^{c=A} > t^h), \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

3.3. Intra-class Query Pixels Sampling Strategy

Instead of sampling only easy and/or hard queries [5], firstly, a dynamic sampling strategy that uses an *adaptive threshold* according to the model's performance on class A is introduced to split all the pixels into an easy query set and a hard query set. Then, *dynamically changed sampling ratios of* the easy and hard queries involved in training following the principle of sampling the most representative pixels. As is shown in Fig. 1, we divide all $p_i^{c=A}$-sorted representations of the class A (the pixel-wise probability of class A, sorted in descending order according to the adaptive threshold derived from the inter-class confidence degree of class A) into easy and hard subsets via the *adaptive threshold* derived from the inter-class confidence degree of the class A itself. It is worth noting that not only the *easy-hard sampling ratio* but also the *threshold* for distinguishing between easy and hard pixels are both *dynamically changed* at different training phases.

3.3.1. Adaptive Threshold for Query Pixels Sampling

We introduce an adaptive threshold for query pixels sampling. The adaptive threshold is formulated in Eq. 3.

$$T^{A} = 1 - \overline{D}_{A}^{c=A}. \tag{3}$$

Here, $\overline{D}_{A}^{c=A}$ is the intra-class-confidence degree of class A. For example, when $\overline{D}_{A}^{c=A}$ is 0.3, the representations with $p_{i}^{c=A} < (1-0.3)$ are marked as hard queries while $p_{i}^{c=A} > (1-0.3)$ are categorized as easy queries. When performance in class A is poor, a high threshold is preferred so that pixels above this high threshold can be grouped into the easy subset, thus sampling pixels from this easy subset to train the model with low performance. When performance is high, a lower threshold is chosen so that the harder pixels can be placed in the hard subset, thus the harder pixels can be used for a model to improve the ability to solve hard problems.

3.3.2. Dynamically Adaptive Sampling Ratios of the Easy and Hard Subsets

The $\overline{D}_{A}^{c=A}$ also assumes the responsibility of adjusting the ratio of easy and hard sample queries. The sampling ratios of easy/hard are set as $(1-\overline{D}_{A}^{c=A})/\overline{D}_{A}^{c=A}$, which aims to sample more easier pixels when the performance of Class A is low and sampling harder pixels when the performance of class A is high. For example, when $\overline{D}_{A}^{c=A}$ is 0.3, that means the predicted ability of class A is weak. During this training phase, easy pixels are more acceptable for optimisation. Thus, the sampling ratios of the easy and hard queries should be 0.7 and 0.3. On the contrary, at a phase where the model is highly generalizable, the well-trained network should focus on the hard samples to further improve model performance.

3.4. Negative Pixels Sampling Strategy

Sampling negative pixels for contrastive learning will help the model learn the class-differentiable details. Classes that are different from the query class are treated as negative classes. The inter-class-confusing degree $\overline{D}_{A}^{c\neq A}$ in Eq. 4 can represent the misclassification probability of all the negative classes.

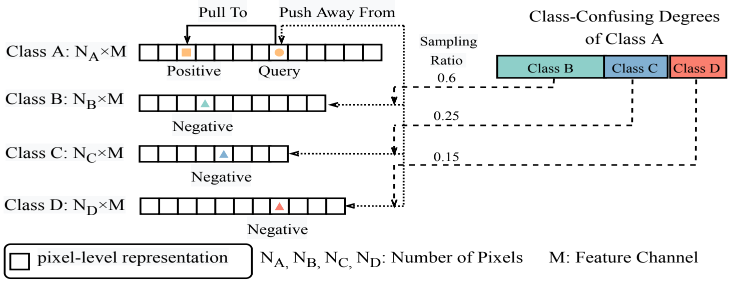

Figure 4: Hard negative class sampling strategy. The representations of all pixels in the image are divided into four groups by class labels, and each group contains the representations of all pixels within a class. The query pixel (a circle) in class A (yellow) is pulled to the positive core (prototype, represented as a square) generated by the easiest pixels in A and pushed away from negatives (triangles), namely the pixels of the other classes (Green, Blue and Red). The sampling rate of samples from the negative classes is determined by the class confusing degree of the query class A. The difference in the lengths of the rectangles of (Class B, C and D) reflects the degree of confusion between class A and other classes.

$$\overline{D}_A^{c \neq A} = avg(\{p_i^{c \neq A} | i \in \mathcal{I}_{\mathcal{A}}\}). \tag{4}$$

To separate the hardest classes from class A, we identify hard negative classes based on *inter-class-confusing degrees* $\overline{D}_A^{c \neq A}$. When selecting negative samples, the negative class with the highest confusion with respect to query class A should be given a higher sampling ratio to further discriminate queries from this negative class. While a class with a small negative class confusion degree is easier to distinguish and thus does not need to sample many negative samples. We use the ratio $[\overline{D}_A^1 : \overline{D}_A^2 :... :\overline{D}_A^c]$ where $c \neq A$, to build negative class sampling ratios. As shown in Fig. 4, Class A is misclassified as Class B with a higher probability of misclassification at 0.6 than Class C (0.25) and Class D (0.15). When a pixel of Class A is used as the query class, the other classes are regarded as negative classes, and the negative samples of the different negative classes should follow the confusing degrees of the class [0.6: 0.25: 0.15] to sample different numbers of pixels in each negative class. If the inter-class-confusing degrees of Class A with the other classes changes as the model's performance

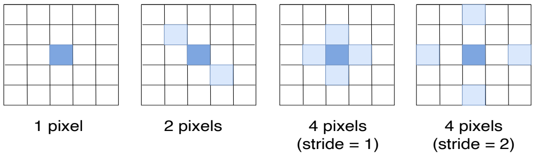

Figure 5: Region spatial layout. The pixel and its neighbors together represent the regional representation of the pixel, and the neighbors include horizontal, vertical and diagonal pixels.

changes, the sampling ratios of the different classes should also be dynamically adjusted.

3.5. Contrastive Enhancement with Spatial Prior

To incorporate spatial dependency into query optimisation by a *spatial enhancement strategy*, the pixel-level queries are enhanced with their neighbours to build region-level queries for contrastive learning. As shown in Fig. 5, we generalize representations of each query pixel with the features of its neighbouring pixels to build region-level representations z^q in the segmentation task, as pixels are spatially correlated.

3.6. Class Confidence and Confusion Guided Dynamic Pixels Sampling Strategy for Contrastive Learning

We use the proposed dynamic sampling strategies in the above sections to sample core, query and negative pixels to construct positive and negative pairs for contrastive learning, intending to select the pixels that are most beneficial for improving the performance. In this section, We will describe in detail how these sampled pixels are used in our contrastive learning framework.

At different training phases, the intra-class confidence degree $\overline{D}_A^{c=A}$ is used to reflect the performance of class A at each training phase and the inter-class confusing degree $\overline{D}_A^{c\neq A}$ is used to reflect misclassification degree of the pixels in class A to the other classes. In the following part of this section, we use θ^+ as the intra-class-confidence degree and θ^- as the inter-class-confusing degree to dynamically change the contrastive sampling strategy. We aim to use different pixel sampling strategies to train a model with different performances. Firstly, we define a joint probability distribution of the easy pixels z_e^q and the hard pixels z_h^q which belong to class A over classes C:

$$P(A|\{z_e^q, \lambda z_e^h\}) = \frac{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+})}{exp(-s^{\theta+}_{\{z_e^q,\lambda z_h^q\},z_A^+}) + \sum_{c\neq A}^{C} exp(-s^{\theta-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-})}, \quad (5)$$

with $s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+} = min\{<\{z_e^q, \lambda z_h^q\}, z_A^+>\}$. $s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+} \in [-1, 1]$ is used to measure the similarity between the query pixels and the core (easiest pixels) of class A. θ^+ is the intra-class confidence degree to control the easy-hard subsets dividing threshold and the sampling ratio λ of intra-class easy and hard query pixels. $s^{\theta^-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-}$ is the query-negative distance. θ^- is the inter-class-confusing degree to control the sampling ratio β of negative pixels of different classes. The CCDC loss of class A for mixed easy and hard queries with a confidence core and hard priority classes negative pixels can be expressed as:

$$\begin{aligned}\mathcal{L}^A_{CCDC} &= -logP(A|\{z_e^q, \lambda z_e^h\}) \\ &= -log\frac{exp(-s^{\theta^+}_{\{z_e^q,\lambda z_h^q\},z_A^+})}{exp(-s^{\theta+}_{\{z_e^q,\lambda z_h^q\},z_A^+}) + \sum_{c\neq A}^{C} exp(-s^{\theta-}_{\{z_e^q,\lambda z_h^q\},\beta z_c^-})}.\end{aligned} \quad (6)$$

Our designed CCDC loss for all the classes can be formulated as:

$$\mathcal{L}_{CCDC} = \frac{1}{C}\sum_{a\in C} -log\frac{exp(-s^{\theta^+}_{\{i_e,\lambda i_h\},a})}{exp(-s^{\theta+}_{\{i_e,\lambda i_h\},a}) + \sum_{c\neq a}^{C} exp(-s^{\theta-}_{i,\beta a})}. \quad (7)$$

The designed CCDC loss measures similarity via the dot product of region-level representations. It is worth mentioning that our CCDC loss is different from the other contrastive loss [5, 6, 36] since the queries, core and negatives in our CCDC, are all dynamically adjusted by different sampling strategies at different training phases. Moreover, our CCDC loss also differs from pixel-only sampling in that we extend the representation to a regional level that contains spatial information. The overall loss is the sum of the cross-entropy loss, the DICE loss, and the CCDC loss in the following way:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{DICE} + \mathcal{L}_{CCDC}. \tag{8}$$

4. Experiment

4.1. Experimental Setup

Dataset. We use three publicly available datasets to evaluate our method. (1) The MS-CMRSeg dataset [17] contains 45 3D cases. Each case is composed of 10-16 2D slices, resulting in a total of 686 slices. Following the setting of the challenge, the LGE modality is applied to our experiments. 10% of the data (5 3D scans) is used for training, and 90% of the data (40 3D scans) is used for testing. (2) CHAOS (Combined (CT-MR) Healthy Abdominal Organ Segmentation) Dataset [18] provides 60 3D DICOM data from 20 patients [T1-DUAL in phase (20 cases), T1-DUAL out phase (20 cases) and T2-SPIR (20 cases)] with four organs ground truth masks: liver, right kidneys (RK), left kidneys (LK) and spleen. The dataset was obtained from a 1.5T Philips MRI, which produces 12-bit DICOM images. The number of slices is between 26 and 50, resulting in a total of 1917 slices. The resolution is approximately 224 × 320. 10% of the data (6 3D scans from 2 patients) is used for training and the remaining 90% of the data (54 3D scans from 18 patients) is used for testing. There is no overlap in patient IDs between the training and testing sets. (3) The ACDC Challenge Dataset [16] consists of 100 patients (200 3D scans), with each case comprising 6-16 2D slices, resulting in a total of 1658 slices. The dataset was acquired using 1.5T and 3T scanners and is accompanied by expert annotations for three structures: the left ventricle (LV), myocardium, and right ventricle (RV). The MRI images from both scanners are adopted in our experiments. There is also no patient overlap between the training and testing sets. The train-test ratio is identical to the setting of the MS-CMRSeg Challenge. In addition to the above experiments, four additional experiments are conducted in the ACDC Challenge Dataset, which use 5%, 10%, 20% and 40% of the dataset for training to further validate the proposed approach.

Implementation details. We use original implemented nnU-Net [15] as our backbone with the same augmentations, such as elastic transformation, rotation, scaling, random crop, scaling, adding Gaussian noise, gaussian blur transformation, brightness multiplicative transformation, contrast augmentation transformation, simulate low-resolution transformation, gamma transformation and mirror transformation. Following the suggestion of nnU-Net, the poly learning rate is also introduced in our training. The experiments on MS-CMRSeg Dataset are conducted with a batch size of 12 and a patch size of 512×512. We use a batch size of 44 and patch size of 224 × 320 on the CHAOS Dataset. For the ACDC Challenge Dataset, the batch size and patch size are 56 and 256×224, respectively. The number of both positive and negative pairs are set to 256. We set the fixed high threshold t^h to 0.97, and the temperature τ in the CCDC loss is 0.5. The ratio of $\mathcal{L}_{CE}$, $\mathcal{L}_{DICE}$ and $\mathcal{L}_{CCDC}$ is 1:1:1. Instead of building a memory bank to store the representations for constructing contrastive pairs, representations are sampled dynamically in a batch during each iteration, enabling less memory consumption. This is also demonstrated in [5], where sampling in a mini batch can achieve results similar to those obtained by methods that use additional memory banks.

4.2. Comparison to the State-of-the-Art

Firstly, we benchmark a state-of-the-art *CNN-based model (nnU-Net)* and a *transformer-based method (nnFormer)* in the ACDC Challenge Dataset, the MS-CMRSeg Dataset and the CHAOS Dataset. The results are reported in Tab. 1, Tab. 2 and Tab. 3. nnU-Net serves as our baseline model to validate our CCDC loss since nnU-Net outperforms nnFormer in our experiment settings on three datasets. Compared to the benchmark (nnU-Net), the DSC sees a rise of 1.84%, 3.12% and 3.68% on the three datasets respectively.

Table 1: Segmentation results (DSC and 95HD) on the ACDC Challenge Dataset.

Method	Pubs.	DSC↑			
		RV	myo	LV	mean
nnFormer [27]	2021	78.55	84.31	91.17	84.68
nnU-Net [15]	Nat. Methods 2021	80.90	84.81	91.11	85.61
Focal [31]	ICCV 2017	79.45	84.54	91.12	85.04
GDL [37]	DLMIA 2017	82.10	86.13	91.70	86.64
TopK [38]	NIPS 2017	81.29	85.01	91.34	85.88
MCC [39]	ISBI 2021	80.55	84.54	90.87	85.32
RegionContrast [36]	ICCV 2021	81.38	85.05	91.66	86.03
ContrastiveSeg [6]	ICCV 2021	**83.96**	84.8	91.33	86.70
ReCo [5]	ICLR 2022	82.26	85.87	91.62	86.58
CCDC (ours)	-	83.49	**86.49**	**92.36**	**87.45**
Method	Pubs.	95HD↓			
		RV	myo	LV	mean
nnFormer [27]	2021	6.26	3.43	4.16	4.61
nnU-Net [15]	Nat. Methods 2021	2.87	1.76	3.16	2.60
Focal [31]	ICCV 2017	3.18	1.63	2.61	2.47
GDL [37]	DLMIA 2017	2.92	**1.58**	2.07	2.19
TopK [38]	NIPS 2017	3.16	1.69	2.55	2.47
MCC [39]	ISBI 2021	2.95	2.08	2.89	2.64
RegionContrast [36]	ICCV 2021	2.44	1.87	**1.68**	2.00
ContrastiveSeg [6]	ICCV 2021	2.38	1.91	2.66	2.32
ReCo [5]	ICLR 2022	3.22	1.66	2.3	2.39
CCDC (ours)	-	**2.35**	1.67	1.70	**1.91**
P-values	-	< 5e-2 (DSC), < 5e-2 (HD95)			

4.2.1. ACDC Challenge Dataset

The results for the three classes show that CCDC consistently outperforms other methods of contrastive learning (ReCo +0.87%, ContrastiveSeg +0.75%

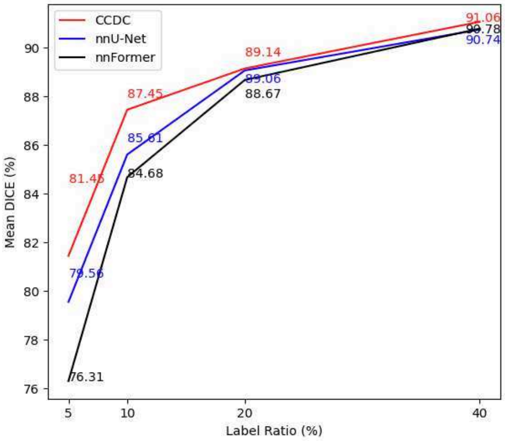

Figure 6: Plot of the mean DSC w.r.t different ratio of the training set over the ACDC Challenge Dataset (5%, 10%, 20%, 40% of the dataset for training), with the blue line representing the result of the baseline: nnU-Net, the black line represents nnFormer, the red line representing our proposed method.

and RegionContrast +1.42%). It may be due to the increased expressiveness of the proposed CCDC model as it relies on optimizing different pixels from easy to hard at different training phases to improve the performance. The average 95HD of all classes is also the smallest, which indicates that performance is improved when CCDC loss is used.

Four additional train-test ratios, 5%, 10%, 20%, 40%, are also evaluated on the ACDC challenge dataset. As shown in Fig. 6, obvious improvements can be observed by comparing the results w/ and w/o CCDC (+1.79%, +1.84%, +0.08%, +0.32%). Notably, our method boosts the baseline model by a significant margin when the amount of the training data is extremely limited. Thus, CCDC demonstrates its potential in facilitating segmentation algorithms in practical applications where labelled data for training is difficult to collect due to concerns about privacy policy and the cost of manual annotations.

4.2.2. MS-CMRSeg Dataset

For MS-CMRSeg dataset, larger patch sizes are adopted, since the height and width of the images in the MS-CMRSeg dataset are much larger compared

Table 2: Segmentation results (DSC and 95HD) on the MS-CMRSeg Dataset.

Method	Pubs.	DSC↑			
		RV	myo	LV	mean
nnFormer[27]	2021	51.96	63.50	70.21	61.89
nnU-Net [15]	Nat. Methods 2021	73.01	74.44	88.95	78.8
Focal [31]	ICCV 2017	74.64	75.26	88.34	79.41
GDL [37]	DLMIA 2017	75.3	73.56	87.57	78.81
TopK [38]	NIPS 2017	72.63	74.22	88.50	78.45
MCC [39]	ISBI 2021	74.08	74.55	87.72	78.78
RegionContrast [36]	ICCV 2021	74.84	76.25	88.51	79.86
ContrastiveSeg [6]	ICCV 2021	75.48	78.44	89.65	81.19
ReCo [5]	ICLR 2022	73.69	75.98	89.23	79.63
CCDC (ours)	-	**76.45**	**79.26**	**90.04**	**81.92**
Method	Pubs.	95HD↓			
		RV	myo	LV	mean
nnFormer[27]	2021	43.89	13.46	13.45	23.60
nnU-Net [15]	Nat. Methods 2021	14.89	9.96	22.79	15.88
Focal [31]	ICCV 2017	15.91	12.6	18.54	15.69
GDL [37]	DLMIA 2017	11.53	9.52	15.12	12.06
TopK [38]	NIPS 2017	13.71	7.57	21.53	14.27
MCC [39]	ISBI 2021	10.69	12.05	24.56	15.76
RegionContrast [36]	ICCV 2021	6.29	5.05	**6.62**	**5.99**
ContrastiveSeg [6]	ICCV 2021	9.54	12.89	24.43	15.62
ReCo [5]	ICLR 2022	6.83	12.18	18.57	12.53
CCDC (ours)	-	**5.60**	**4.31**	13.79	7.90
P-values	-	< 5e-2 (DSC), < 5e-2 (HD95)			

to the ACDC Challenge Dataset (from 256×224 to 512×512). As shown in Tab. 2, CCDC substantially exceeds the state-of-the-art methods by [36, 6, 5], in terms of class-specific DSC scores, our method outperforms the state-of-the-art methods in all classes and the average DSC is significantly improved (2.06% improvement with RegionContrast, 0.73% with ContrastiveSeg and 2.29% with ReCo on DSC). This demonstrates that CCDC is consistently beneficial for the baseline model, whether the training data are small images or large images with rich details. We also show class-specific results in the 95HD metrics. The 95HD is the smallest in the class of RV and myo and the average 95HD of CCDC also ranks second best.

4.2.3. CHAOS Dataset

We evaluated our method using the CHAOS dataset (four classes segmentation tasks). Tab. 3 shows the results of the segmentation for the liver, RK, LK and spleen. We can see that all contrastive learning methods [36, 6, 5] show improvements over baseline on DSC, demonstrating the benefits of contrastive learning. Our CCDC learning with a dynamic learning strategy consistently keeps improving on the average DSC score. These results demonstrate the effectiveness of our dynamic learning strategy with extremely limited data (6 images from 2 patients only) and validate that our dynamic sampling strategy can be applied to a wider range of multi-class segmentation tasks. The 95HD of the CCDC(ours) algorithm is not optimal, and the DSC of the liver and spleen are not optimal. We have conducted further analysis and found that it may be due to the dataset we used in the experiment contains more classes and many images that contain a high degree of variability in terms of size, shape, and texture of the structures of interest. They together make it challenging to achieve optimal results for all classes in multi-class tasks, where the 95HD metric might not be as relevant as the DSC. This may be a result of the class imbalance or overlapping features among classes. Despite these challenges, our overall performance on the DSC metric has still improved.

Table 3: Segmentation results (DSC and 95HD) on the CHAOS Dataset.

Method	Pubs.	DSC↑				
		liver	RK	LK	spleen	mean
nnFormer[27]	2021	70.88	58.41	56.36	62.66	62.08
nnU-Net [15]	Nat. Methods 2021	73.79	57.09	63.27	60.27	63.61
Focal [31]	ICCV 2017	76.31	59.28	67.43	61.95	66.24
GDL [37]	DLMIA 2017	74.8	54.65	65.17	61.26	63.97
TopK [38]	NIPS 2017	75.85	55.14	68.63	**62.68**	65.57
MCC [39]	ISBI 2021	75.63	56.08	63.88	61.09	64.17
RegionContrast [36]	ICCV 2021	74.45	57.15	68.58	57.40	64.39
ContrastiveSeg [6]	ICCV 2021	**77.52**	56.59	66.11	57.36	64.40
ReCo [5]	ICLR 2022	75.44	61.85	67.66	61.05	66.50
CCDC (ours)	-	75.03	**62.43**	**70.81**	60.89	**67.29**
Method	Pubs.	95HD↓				
		liver	RK	LK	spleen	mean
nnFormer[27]	2021	27.58	29.24	30.35	28.99	29.04
nnU-Net [15]	Nat. Methods 2021	17.80	37.93	30.16	21.19	26.77
Focal [31]	ICCV 2017	13.86	23.05	19.6	16.85	18.34
GDL [37]	DLMIA 2017	14.16	20.30	30.68	25.86	22.75
TopK [38]	NIPS 2017	14.49	14.62	9.90	36.77	18.95
MCC [39]	ISBI 2021	13.27	20.31	25.28	**16.04**	18.72
RegionContrast [36]	ICCV 2021	14.75	**12.74**	**7.34**	25.01	**14.96**
ContrastiveSeg [6]	ICCV 2021	**12.45**	19.21	22.03	33.80	21.87
ReCo [5]	ICLR 2022	16.31	20.90	24.94	23.17	21.33
CCDC (ours)	-	14.51	20.23	23.97	19.64	19.59
P-values	-	< 5e-2 (DSC), < 5e-2 (HD95)				

Table 4: Dynamic query pixels sampling strategy on ACDC Challenge Dataset.

Method	Dynamic Query Sampling		DSC↑			
	Adaptive T	Adjusted Ratio	RV	myo	LV	mean
W/O Contrastive Learning			80.90	84.81	91.11	85.61
Pixel-level	-	-	82.13	86.15	91.81	86.70
	√	-	82.39	85.92	91.78	86.70
	-	√	82.60	85.80	92.15	86.85
	√	√	83.84	85.82	91.64	**87.10**
Region-level	-	-	80.85	85.76	91.72	86.11
	√	-	82.21	86.86	92.38	87.15
	-	√	82.68	85.59	91.84	86.71
	√	√	83.49	86.49	92.36	**87.45**

4.3. Ablation Studies

4.3.1. The Impact of Dynamic Query Sampling Strategy with Pixel-Level and Region-Level Representations

Tab. 4 shows the ablation study with and without the dynamic contrastive learning in easy and hard subsets which are separated by an adaptive threshold (adaptive T) with the adjusted ratio for sampling easy to hard queries on the ACDC challenge dataset.

We examined our proposed dynamic contrastive learning on pixel-level representations and region-level representations separately. It is worth noting that despite not using dynamic sampling strategies, contrastive learning improves performance with and without the representation enhancement strategy. More notably, when using a fixed threshold to build easy and hard sets, the introduction of dynamic contrastive learning to adjust easy-hard ratios improves the results by using the pixel-level or region-level representations (Pixel-level: 86.7% to 86.85%, Region-level: 86.11% to 86.71%). Fixing the ratios of sampling easy and hard queries and using the changed threshold to select easy and hard sets to obtain a similar average score (86.7%) in pixel-level contrastive learning, but

1.04% (86.11% to 87.15%) improvement at the region level. This shows that our the dynamic sampling strategy can achieve some performance improvement regardless of adjusting the threshold of the easy and hard sets or their sampling ratio. When using the adaptive threshold and the adjusted sampling ratio together to separate easy and hard sets, the performance is further improved, suggesting that the dynamic approach can capture the pixels that are most in need of optimisation at different training phases to improve segmentation performance.

Table 5: Query sampling from different intra-class subsets on ACDC Challenge Dataset.

Method	Sampling Strategy	DSC↑			
		RV	myo	LV	mean
W/O Contrastive Learning		80.90	84.81	91.11	85.61
Pixel-level	Hard	80.29	85.77	90.86	85.64
	Easy	81.87	85.66	92.17	86.57
	Easy to Hard	81.67	85.27	91.47	86.14
	Dynamic Sampling	83.84	85.82	91.64	**87.10**
Region-level	Hard	82.34	86.14	92.01	86.83
	Easy	82.22	86.95	92.38	87.18
	Easy to Hard	83.87	86.12	91.86	87.28
	Dynamic Sampling	83.49	86.49	92.36	**87.45**

4.3.2. The Impact of Query Sampling from Different Subsets

Here, we focus on different strategies for sampling query pixels: 1) sampling queries in the easy set; 2) sampling queries in the hard set; 3) sampling queries from easy to hard; 4) dynamically changing strategies for queries sampling. As described in the methodology section, we do not use fixed thresholds to identify easy and hard sets, but instead use an adaptive threshold to separate the easy and hard queries to be sampled based on how well the network is trained during different phases. The pixel-level representations and the region-level representations are both used for the sampling strategy comparison. Tab. 5 shows the results of the four sampling strategies on the ACDC challenge dataset. The four aforementioned sampling strategies yield better results than the benchmark. The results of the easy sampling strategy consistently outperform those of the hard-only sampling (Pixel-level: 85.64% to 86.57%, Region-level: 86.83% to 87.18%). The proposed dynamic sampling strategy further improves the performance (Pixel-level: +0.53%, Region-level: +0.27%). It verified that using only hard pixels for training may result in the network not converging consistently. Training with only easy pixels can achieve better performance than only hard pixels, but the performance is also lower than dynamic sampling strategy. This is because using only easy pixels does not provide enough gradient descent for optimization, which may influence performance improvement. It can be seen that the performance of the easy-to-hard sampling strategy with pixel-level representations decreased when compared with only using easy pixels, this may be because easy-to-hard sampling does not measure the current performance of the model and may not necessarily provide the pixels needed by the current model. However, the performance of the easy-to-hard strategy with region-level representations is better than using only easy or hard pixels, but not as good as our dynamic sampling strategy. As our dynamic sampling strategy can dynamically change the sampling strategy based on the performance of the model and can select the most representative pixels at different training stages. The result reveals that our method can, not only improve the basic generalization ability when the performance is low, but also enhance the hard problem-solving ability when the performance is strong, suggesting that the approach is beneficial for seeking better pixel-level segmentation applications.

4.3.3. The Impact of Hard Negative Class Mining Strategy

Tab. 6 shows the performance improvement of the ACDC Challenge Dataset for different negative class sampling strategies. Despite using different sampling strategies, we find that CCDC loss has a consistent improvement of 1%-2%. Uniform sampling results in 0.29% improvement compared to random sampling. This may be due to the class imbalance problem, if a class in a mini-batch has

Table 6: Hard negative class mining strategy on ACDC Challenge Dataset

Method	DSC↑			
	RV	myo	LV	mean
Baseline	80.9	84.81	91.11	85.61
Random Sampling	81.91	86.29	92.15	86.78
Uniform Sampling	84.12	84.78	92.30	87.07
Dynamic Sampling	83.49	86.49	92.36	**87.45**

a small number of pixels, and this class is also hard to distinguish, random sampling is not sufficient when sampling from this hard class for training, resulting in missing samples and low performance on this hard class. This can also be demonstrated by the results of the hard but fewer pixels RV class (DSC of Random: 81.91% and DSC of UniForm 84.12%). Following our class confusing degree-guided hard-negative class mining strategy, negative samples can be sampled from each class with different ratios, which is more beneficial in sampling negative samples thus obtaining significant performance improvement.

4.3.4. The Impact of Region Enhancement Strategy

Table 7: Region enhancement strategy on the ACDC Challenge Dataset.

Number of Pixels	DSC↑			
	RV	myo	LV	mean
1	83.84	85.82	91.64	87.10
2	83.73	86.30	91.36	87.13
4 (Stride=2)	83.05	86.36	92.60	87.34
4 (Stride=1)	83.49	86.49	92.36	**87.45**

We evaluated the effectiveness of the region enhancement strategy. Different region enhancement schemes are adopted and shown in Fig. 5. Specifically, using different numbers of pixels and strides in diagonal, horizontal and vertical directions to build regions can achieve different improvements. As shown in Tab. 7, the region enhancement strategy can achieve a consistent improvement over the benchmark, and it is observed that the performance is improved with the increase of the number of neighbouring pixels. The closer the neighboring pixels are to the center pixel, the further the performance can be improved. This is because pixels in boundary regions are often more difficult to distinguish and require stronger local location information for fine differentiation. In contrast, our representation enhancement strategy introduces local location information from the neighbouring pixels for contrastive learning, thus providing a supplement to determine the cross-class boundaries.

4.3.5. The Impact of the Fixed High Threshold to Choose Confidence Core

Table 8: Different fixed high threshold on ACDC Challenge Dataset.

High Threshold	DSC↑			
	RV	myo	LV	mean
W/O Dynamic	80.90	84.81	91.11	85.61
0.50	83.27	86.00	92.39	87.22
0.75	82.78	85.50	92.46	86.91
0.90	82.97	86.50	92.53	87.33
0.97	83.49	86.49	92.36	**87.45**

We evaluated the impact of a fixed high threshold used to select the confidence core. In Tab. 8, it can be observed that slightly better performance can be obtained with a relatively large threshold (0.97 and 0.9). When the confidence threshold drops to 0.75 and 0.5, compared with using the thresholds over 0.9, performance declined to a 87.22% and 86.91%. However, compared to the baseline, the performance is still greatly improved. We can also observe that our method is not highly dependent on fixed high-threshold hyperparameters for confidence pixel selection. This is because a fixed threshold is needed to ensure that the selected pixel can be predicted as the correct class. Thus, the pixels above this threshold can form a class-specific prototype to guide the learning of other low-confidence pixels within the class.

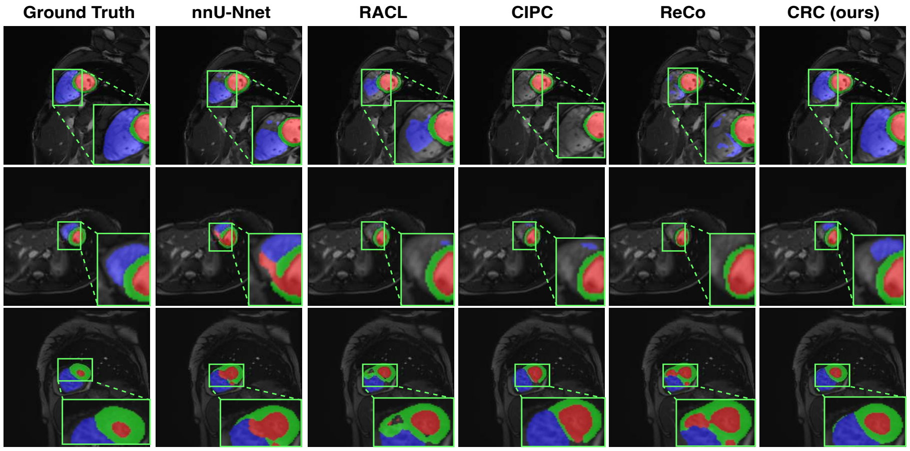

Figure 7: Typical segmentation result comparing different approaches on the ACDC Challenge Dataset. Blue, green and red denote the classes of RV, myocardium, and LV, respectively.

5. Visualization

Fig. 7 shows the segmentation results using a model trained with 10% data on the ACDA dataset when compared with the state-of-the-art contrastive learning methods [36, 6, 5]. The prediction results of the RV class in the first row and the myocardium class in the second and third rows show that our method achieves significant improvements in the identification of boundaries and the segmentation of hard classes. This is attributed to the dynamic learning strategy which identifies the core representations of a class and pulls different hardness pixels to the core along with the training, Furthermore, negative class mining also identifies the hard negative classes, avoiding a large number of pixels being misclassified to the hard negative classes. The representation enhancement strategy introduces local spatial information, further refining the determining of boundaries and reducing the confusion between different adjacent classes.

6. Conclusion and Discussion

In this paper, we present a dynamic contrastive learning framework for medical image segmentation that leverages class-confidence and confusion. The framework features a novel loss function, the CCDC loss, which dynamically samples contrastive pixels based on the model's performance. Our approach selects the most expressive pixels as positive and negative pairs during different training phases and employs a hard negative class mining strategy to enhance effectiveness. The results show that our method outperforms the state-of-the-art on three challenging datasets and has potential for other multi-class tasks such as classification and detection. While our approach has achieved significant progress, we acknowledge the need to address the dynamic adjustment of easy/hard boundaries in situations where the dataset is challenging. To enhance robustness, we suggest potential future research directions, including implementing a stopping criterion. Future work also includes exploring the method's applicability to other domains or tasks, scalability, efficiency, robustness, and generalization. Another avenue for future research is to investigate combining our method with other techniques to further improve performance.

References

[1] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[2] Z. Cao, W. Wang, L. Huo, S. Niu, Unsupervised class-to-class translation for domain variations, Pattern Recognition 138 (2023) 109346.

[3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[4] A. Ben Saad, K. Prokopetc, J. Kherroubi, A. Davy, A. Courtois, G. Facciolo, Improving pixel-level contrastive learning by leveraging exogenous depth information, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2380–2389.

[5] S. Liu, S. Zhi, E. Johns, A. J. Davison, Bootstrapping semantic segmentation with regional contrast, in: International Conference on Learning Representations, 2022.

[6] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7303–7313.

[7] T. Sun, C. Lu, T. Zhang, H. Ling, Safe self-refinement for transformer-based domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7191–7200.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE transactions on pattern analysis and machine intelligence 32 (9) (2010) 1627–1645.

[9] G. Hacohen, D. Weinshall, On the power of curriculum learning in training deep networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 2535–2544.

[10] A. Graves, M. G. Bellemare, J. Menick, R. Munos, K. Kavukcuoglu, Automated curriculum learning for neural networks, in: international conference on machine learning, PMLR, 2017, pp. 1311–1320.

[11] J. L. Elman, Learning and development in neural networks: The importance of starting small, Cognition 48 (1) (1993) 71–99.

[12] Y. Wang, W. Gan, J. Yang, W. Wu, J. Yan, Dynamic curriculum learning for imbalanced data classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5017–5026.

[13] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769.

[14] X. Zhang, Q. Guo, Y. Sun, H. Liu, G. Wang, Q. Su, C. Zhang, Patch-based fuzzy clustering for image segmentation, Soft Computing 23 (9) (2019) 3081–3093.

[15] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2) (2021) 203–211.

[16] X. Zhuang, J. Xu, X. Luo, C. Chen, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar, et al., Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge, Medical Image Analysis 81 (2022) 102528.

[17] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?, IEEE transactions on medical imaging 37 (11) (2018) 2514–2525.

[18] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al., Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation, Medical Image Analysis 69 (2021) 101950.

[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[20] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[21] K. Wang, X. Zhang, Y. Lu, W. Zhang, S. Huang, D. Yang, Gsal: Geometric structure adversarial learning for robust medical image segmentation, Pattern Recognition 140 (2023) 109596.

[22] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, D. Yang, Eanet: Iterative edge attention network for medical image segmentation, Pattern Recognition 127 (2022) 108636.

[23] J. Chen, J. Zhang, K. Debattista, J. Han, Semi-supervised unpaired medical image segmentation through task-affinity consistency, IEEE Transactions on Medical Imaging.

[24] J. Chen, H. Li, J. Zhang, B. Menze, Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac mr images segmentation, in: Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10, Springer, 2020, pp. 317–325.

[25] J. Chen, W. Li, H. Li, J. Zhang, Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer, 2020, pp. 187–196.

[26] Q. Yan, S. Liu, S. Xu, C. Dong, Z. Li, J. Q. Shi, Y. Zhang, D. Dai, 3d medical image segmentation using parallel transformers, Pattern Recognition 138 (2023) 109432.

[27] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: Interleaved transformer for volumetric segmentation, arXiv preprint arXiv:2109.03201.

[28] Z. Liu, F. Wu, Y. Wang, M. Yang, X. Pan, Fedcl: Federated contrastive learning for multi-center medical image classification, Pattern Recognition (2023) 109739.

[29] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard negative mixing for contrastive learning, Advances in Neural Information Processing Systems 33 (2020) 21798–21809.

[30] K. Yuan, G. Schaefer, Y.-K. Lai, Y. Wang, X. Liu, L. Guan, H. Fang, A multi-strategy contrastive learning framework for weakly supervised semantic segmentation, Pattern Recognition 137 (2023) 109298.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[32] M. S. Hossain, J. M. Betts, A. P. Paplinski, Dual focal loss to address class imbalance in semantic segmentation, Neurocomputing 462 (2021) 69–87.

[33] X. Wang, Y. Chen, W. Zhu, A survey on curriculum learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9) (2021) 4555–4576.

[34] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. A. González Ballester, G. Piella, Medical-based deep curriculum learning for improved fracture classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 694–702.

[35] X. Yao, X. Feng, J. Han, G. Cheng, L. Guo, Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning, IEEE Transactions on Geoscience and Remote Sensing 59 (1) (2020) 675–685.

[36] H. Hu, J. Cui, L. Wang, Region-aware contrastive learning for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16291–16301.

[37] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[38] Y. Fan, S. Lyu, Y. Ying, B. Hu, Learning with average top-k loss, Advances in neural information processing systems 30.

[39] K. Abhishek, G. Hamarneh, Matthews correlation coefficient loss for deep convolutional networks: Application to skin lesion segmentation, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), IEEE, 2021, pp. 225–229.