Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 9960199, 9 pages
https://doi.org/10.1155/2021/9960199

Research Article

LCC-Net: A Lightweight Cross-Consistency Network for Semisupervised Cardiac MR Image Segmentation

Lai Song,[1] **Jiajin Yi**,[1] **and Jialin Peng**[1,2]

[1]*College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[2]*Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, China*

Correspondence should be addressed to Jialin Peng; 2004pjl@163.com

Received 23 March 2021; Revised 22 April 2021; Accepted 29 April 2021; Published 17 May 2021

Academic Editor: Yuankai Huo

Semantic segmentation plays a crucial role in cardiac magnetic resonance (MR) image analysis. Although supervised deep learning methods have made significant performance improvements, they highly rely on a large amount of pixel-wise annotated data, which are often unavailable in clinical practices. Besides, top-performing methods usually have a vast number of parameters, which result in high computation complexity for model training and testing. This study addresses cardiac image segmentation in scenarios where few labeled data are available with a lightweight cross-consistency network named LCC-Net. Specifically, to reduce the risk of overfitting on small labeled datasets, we substitute computationally intensive standard convolutions with a lightweight module. To leverage plenty of unlabeled data, we introduce extreme consistency learning, which enforces equivariant constraints on the predictions of different perturbed versions of the input image. Cutting and mixing different training images, as an extreme perturbation on both the labeled and unlabeled data, are utilized to enhance the robust representation learning. Extensive comparisons demonstrate that the proposed model shows promising performance with high annotation- and computation-efficiency. With only two annotated subjects for model training, the LCC-Net obtains a performance gain of 14.4% in the mean Dice over the baseline U-Net trained from scratch.

1. Introduction

Medical image analysis plays an increasingly important role in routine clinical work. Magnetic resonance imaging (MRI) is a noninvasive technique for investigating cardiac structures, thus widely used in clinical diagnosis and treatment. Segmentation of the left ventricle (LV), right ventricle (RV), and the myocardium (MYO) from cardiac MR images can provide crucial diagnostic parameters about the cardiac. Recently, convolutional neural networks (CNNs), mostly fully convolutional networks (FCNs) [1, 2], have made substantial progress for cardiac image segmentation [3]. However, the current supervised-learning models rely heavily on a large amount of manually labeled data for model training to achieve competitive performance. Unfortunately, manually labeling cardiac MR images is time-consuming and labor-intensive and requires strong domain knowledge from experts. Moreover, most of the top-performing methods are deep and wide convolutional neural networks involving a massive number of training parameters, which not only increases the chance of overfitting but also hinders their applications in clinical routines. To address the above problems, we introduce a lightweight deep network for semisupervised segmentation of cardiac images. Our model is trained only on a few labeled subjects and a more considerable number of unlabeled subjects.

There are generally two paradigms to make use of unlabeled data. The first one is unsupervised or self-supervised pretraining, followed by fine-tuning on a small set of labeled data. The second paradigm is to jointly use the labeled data and unlabeled data through pseudo labeling [4] or consistency regularization [5–8]. Since there is an obvious gap between the objectives of the unsupervised pretraining and the downstream segmentation, the effect of unsupervised pretraining is not always significant. In this study, we follow the second paradigm and make use of the unlabeled data by enforcing consistency regularization on the supervised model, aiming to improve the generalization ability of the

supervised trained model and reduce the risk of overfitting. Consistency regularization encourages the segmentation prediction to be consistent on the unlabeled examples under different data perturbations or among different models. We follow the studies in [6, 9, 10] and enforce consistency among different models' predictions. Both strong and weak perturbations are applied.

In this study, we propose a lightweight network, LCC-Net, for semisupervised segmentation of cardiac MR images based on consistency training cross models. To be specific, our model, as shown in Figure 1, consists of one shared encoder and three separate decoders: one decoder for supervised learning and the other two decoders for unsupervised consistency learning. Following a similar strategy as in [6], different perturbations are injected on the two unsupervised decoders. We enforce consistency between the predictions of the supervised decoder and unsupervised decoders to make the learned model less sensitive to the extra perturbation. To further improve model robustness and reduce the risk of overfitting, we augment the input data, both the labeled and unlabeled data, with extreme perturbations realizing significant gains. While the previous semisupervised models suffer from a massive scale of parameters and high computational complexity, we lighten our model with the lightweight Ghost module introduced in [11]. Moreover, we validate the proposed method on the ACDC [12] dataset.

The rest of the paper is organized as follows. In Section 2, we briefly review the related work. Section 3 presents the proposed method, which is evaluated on challenging cardiac segmentation tasks in Section 4. Section 5 concludes this study.

2. Related Work

2.1. Cardiac MR Image Segmentation Methods. For cardiac MR image segmentation, Painchaud et al. [13] presented a postprocessing VAE [14], which converts anatomically invalid cardiac shapes into close but correct shapes for introducing strong anatomical guarantees into the network. Khened et al. [15] proposed Densely Connected Fully Convolutional Network (DFCN), which is based on DenseNets [16]. Yang et al. [17] proposed a general and fully automatic solution to concurrently segment three important ventricular structures, starting from 3D Fully Convolutional Network (3D FCN). Simantiris and Tziritas [18] proposed a different Dilated CNN structure that incorporating domain-specific constraints. Isensee et al. [19] combined 2D U-Net and 3D U-Net, obtaining the best performance on the ACDC dataset. However, due to the combination of two different models, the numbers of model params is enormous. All these methods base on supervised learning proposed a series of efficient methods from different perspectives. When it comes to semisupervised cardiac MR image segmentation methods, there are still limitations for obtaining remarkable performance because cardiac MR image segmentation is a particular issue, including unique data distribution and difficult segmentation tasks.

2.2. Semisupervised Learning Methods. As for general semisupervised learning, many methods are proposed to reduce the burden of pixel-wise manual annotations for images, such as pseudo labeling [1], graph-based methods [20, 21], and entropy minimization [5]. Besides, mean-teacher [9] is another notable paradigm for semisupervised learning, which could be used in medical image segmentation. The mean-teacher model has two subnetworks: the teacher network and the student network, and learn cross-consistency from unlabeled data by exerting different perturbances on the two subnetworks. Yu et al. [22] proposed the uncertainty-aware mean teacher (UA-MT) framework, learning from the meaningful and reliable targets by exploiting the uncertainty information. Adversarial learning [23] methods are aimed at matching labeled and unlabeled images and improving testing time performance. Hung et al. [24] proposed a novel method in semisupervised semantic segmentation by introducing adversarial learning. Nie et al. [25] proposed attention-based semisupervised deep networks (ASDNet), where they integrated adversarial learning by a confidence network. Virtual Adversarial Training (VAT) [26] utilizes adversarial learning from a novel perspective and alters the model's predictions the most by approximating the perturbations. Laine and Aila [10] introduced consistency regularization into semisupervised learning, including π-model [10] and temporal ensembling method [10]. Bortsova et al. [27] proposed a novel semisupervised method that learns to predict segmentations consistent under a given class of transformations on both labeled and unlabeled images. The above methods enforce the consistency between predictions and provide critical data information to the supervised trained model. Besides, a series of strong data augmentation methods are proposed for overcoming the limitation of labeled training data, such as MixUp [28], CutMix [29], and Mosaic [30]. CowMix [31] starts from MixUp and enforces the consistency between the mixed outputs and the prediction over the mixed inputs. All the above data augmentation methods have made efforts to semisupervised learning by increasing training data diversity.

2.3. Lightweight Deep Networks. Current existing lightweight methods for networks can be divided into model compression and lightweight architecture design. We mainly review methods designing lightweight architectures, which are more related to our study. The increasing need to deploy deep models on computationally limited platforms and process large-scale data encourages lightweight architecture design. A series of lightweight convolutional modules have been proposed to balance the model performance and computational complexity. In particular, depth-wise convolution [32] and group convolution [33, 34] have gained much attention and have been building blocks for many lightweight architectures. MobileNet [35] used depth-wise separable convolution [32], a combination of depth-wise convolution and point-wise convolution, to build a lightweight model. ShuffleNet [36] is presented with point-wise group convolution and channel shuffle, which improves the information flow exchange between channel groups. Recently, Han et al. [11] proposed GhostNet with a novel Ghost module, which utilizes group convolution to further explore correlation and redundancy between feature maps. The GhostNet has shown higher

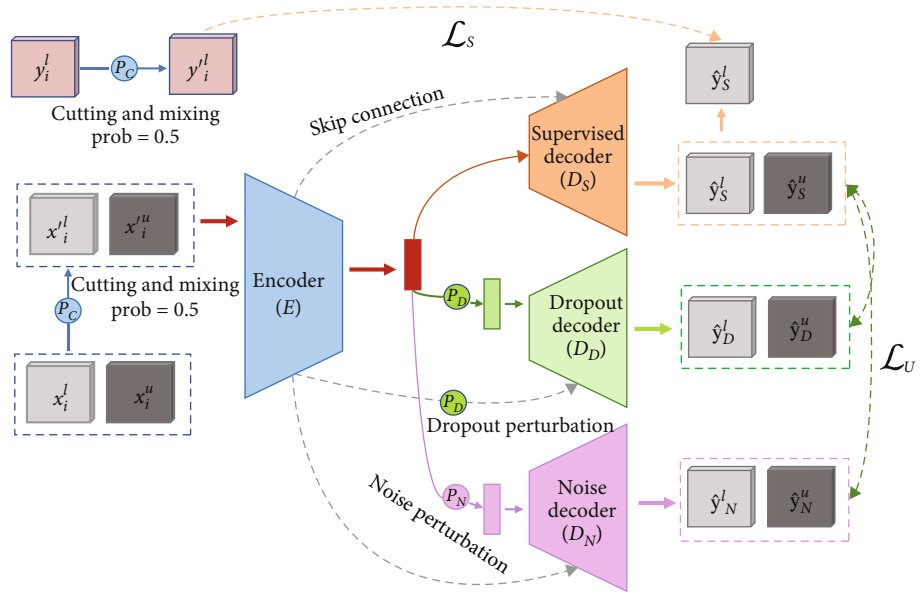


FIGURE 1: Overview of our LCC-Net for semisupervised segmentation. The network contains one supervised decoder and two unsupervised decoders. While the supervised decoder is trained with the labeled data, the two auxiliary decoders are trained with both labeled data and unlabeled data using unsupervised consistency losses. We inject dropout perturbation and noise perturbation in the feature space and inject cutting and mixing perturbation in the input image space.

recognition performance in natural images but has not been applied in medical image segmentation tasks.

3. Methods

3.1. Problem Formulation. We aim to develop a deep network model for semantic segmentation of cardiac MR images with only a few annotated subjects and a larger set of unlabeled subjects. We segment cardiac MR sequences in a slice-by-slice manner. Assume $\mathcal{D}_l = \{X^l, Y^l\}$ denote the labeled data, in which $X^l = \{x_1^l, x_2^l, \cdots, x_n^l\}$ contains n image slices, and $Y^l = \{y_1^l, y_2^l, \cdots, y_n^l\}$ is ground truth. $\mathcal{D}_u = \{x_1^u, \cdots, x_m^u\}$ denotes m unlabeled examples. Usually, the number of unlabeled slices is much larger than labeled ones ($m \gg n$). Making better use of unlabeled data is a critical part of training a semisupervised segmentation network with better generalization ability on unseen data.

An overview of the proposed LCC-Net is demonstrated in Figure 1. We leverage the unlabeled data during supervised segmentation model learning and encourage segmentation consistency on all data under different perturbations with two unsupervised consistency losses. Our segmentation network is in encoder-decoder architecture. Specifically, the LCC-Net contains a shared encoder E and three independent decoders: the supervised decoder D_S, the dropout decoder D_D, and the noise decoder D_N. The encoder E and the decoder D_S constitute the segmentation network $f_S = D_S \circ E$. While the supervised decoder D_S is trained with the labeled data, the two auxiliary decoders are trained with both labeled data and unlabeled data.

We inject perturbations in both the feature space, i.e., the output of the feature encoder E and the input image space.

(i) For perturbations in the feature space, we use two perturbations: dropout perturbation P_D and noise perturbation P_N. The dropout decoder D_D and noise decoder D_N are used to decode the two perturbed versions of features, respectively. We enforce the consistency of predictions between the supervised decoder D_S and the auxiliary decoders D_D and D_N with unsupervised consistency losses. These two auxiliary decoders together with the encoder and feature perturbations constitute the two auxiliary networks $f_D = D_D \circ P_D \circ E$ and $f_N = D_N \circ P_N \circ E$. In the experiments, we use Gaussian noises for the noise perturbation P_N and 10%-40% spatial random dropout for the dropout perturbation P_D

(ii) For perturbations in the image space, we use a stronger perturbation P_C to achieve better model robustness. Specifically, we exploit an adapted version of the Cutmix [29], as illustrated in Figure 2. Given two input images, we first split the images into four blocks of equal size. Then, we randomly exchange one or two blocks on the corresponding positions between the two images. When the two input images are labeled, the corresponding operations are also applied to their label images

We apply the cutting and mixing perturbation on both the labeled data and unlabeled data as a data augmentation to the original data. In addition to the (augmented) unlabeled data, we also feed the perturbed labeled data to the auxiliary networks and enforce cross-model consistency.

3.2. Supervised Training on Few Labeled Data. The segmentation network $f_S = D_S \circ E$ is trained with the (augmented)

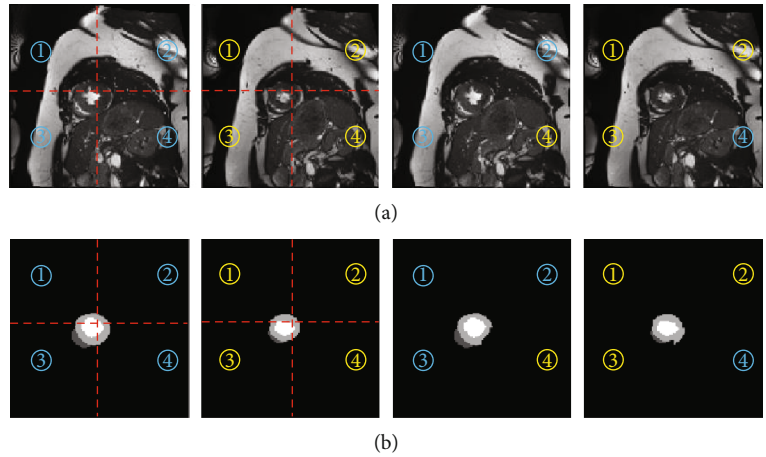


Figure 2: Illustration of the input space perturbation P_C used in our study. (a) The cardiac MR images. From left to right: original image A, original image B, the perturbed image A', and the perturbed image B'. (b) Their corresponding label images. We evenly split an image into four blocks and then randomly exchange one or two blocks on the corresponding positions. The same operations are applied on their ground truth label images, as in (b). Best view in color.

labeled data using a cross-entropy- (CE-) based supervised loss. We also denote the augmented labeled data as $\mathscr{D}_l \cup \mathscr{D}_l'$, where $\mathscr{D}_l'$ is generated by perturbing the images in $\mathscr{D}_l$ using cutting and mixing P_C.

$$\mathscr{L}_S = \frac{1}{\left|\mathscr{D}_l \cup \mathscr{D}_l'\right|} \sum_{(x_i,y_i)\in \mathscr{D}_l \cup \mathscr{D}_l'} l_{\mathrm{CE}}(x_i, y_i), \tag{1}$$

where l_{CE} denotes the cross-entropy loss. The input image x_i^l can be the original image and its perturbed version.

3.3. Unsupervised Cross-Consistency Training. As mentioned above, we enforce cross-model consistency between the predictions of the supervised decoder D_S and the auxiliary decoders D_D and D_N with an unsupervised consistency loss. We denote the augmented unlabeled data as $\mathscr{D}_u \cup \mathscr{D}_u'$, where $\mathscr{D}_u'$ is generated by perturbing the images in $\mathscr{D}_u$ using cutting and mixing P_C. The two auxiliary networks f_D and f_N take both the (augmented) unlabeled data $\mathscr{D}_u \cup \mathscr{D}_u'$ and the perturbed labeled data $\mathscr{D}_l'$. The two auxiliary networks are trained with the following loss.

$$\mathscr{L}_U = \frac{1}{\left|\mathscr{D}_u \cup \mathscr{D}_u' \cup \mathscr{D}_l'\right|} \sum_{(x_i,y_i)\in \mathscr{D}_u \cup \mathscr{D}_u' \cup \mathscr{D}_l'} \cdot \left[\mathbf{d}(f_S(x_i), f_D(x_i)) + \mathbf{d}(f_S(x_i), f_N(x_i))\right], \tag{2}$$

where the distance measure $\mathbf{d}$ is used to measure the consistency of the predictions by different models. In the experiments, we use mean squared error (MSE) as the distance measure.

3.4. The Overall Loss. By integrating the supervised loss and unsupervised loss, the loss of our LCC-Net reads

$$\mathscr{L} = \mathscr{L}_S + \lambda \mathscr{L}_U, \tag{3}$$

in which λ is the trade-off parameter. In the experiments, we choose an exp-schedule function as follows:

$$\lambda(\mathrm{epoch}) = \min\left(\lambda_{\max}, \lambda_{\max} \times e^{2\cdot(\mathrm{epoch}/\mathrm{stop}-1)}\right), \tag{4}$$

in which epoch as current training epoch, stop is the max number of epochs to stop increasing λ, and $\lambda_{\max}$ is an upbound of λ.

3.5. The Backbone of the LCC-Net. To avoid overfitting on the small labeled data, we introduce a lightweight segmentation U-Net (L-Unet) as our backbone network, which is demonstrated in Figure 3. The network is an encoder-decoder with skip-connections between the corresponding layers of the encoder and decoder. To lighten the U-Net, we upgrade the U-Net with lightweight convolutional modules. More precisely, we replace the standard convolutions in U-Net with the Ghost module [11], which involves much fewer parameters and computation costs. The Ghost module is shown in Figure 4. For a feature map $F \in \mathbb{R}^{a\times h\times w}$, in which a is the channel number, and $h \times w$ is the spatial size, we first compress F into $F' \in \mathbb{R}^{(b/s)\times h\times w}$ by using a standard 3×3 convolution, where b is the channel number of the final output, and s is the ratio. Then, we apply $s(=4)$ linear transformations, including one identity transform, on each channel of F' separately to generate s groups of new features, each of which contains b/s feature maps. The linear transformations are achieved with 3×3 convolutions. At last, we concatenate all the generate feature maps and obtain the final output $\widehat{F} \in \mathbb{R}^{b\times h\times w}$. Note that the computation costs of the linear transformations are much lower than standard convolutions.

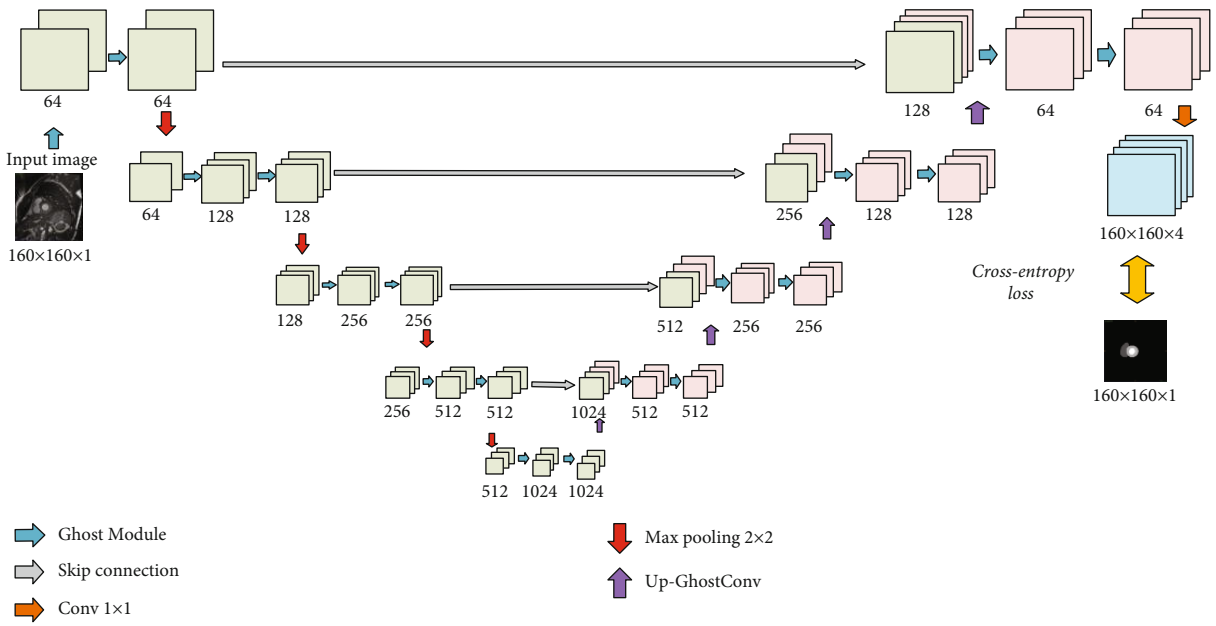


Figure 3: The backbone network of our proposed model, L-Unet. Instead of standard 2D convolutions, the L-Unet uses the Ghost module [11] as the basic building block.

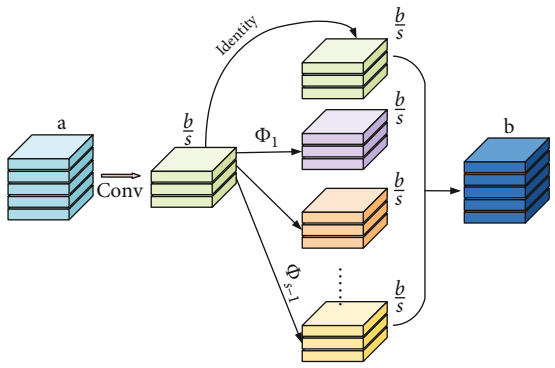


Figure 4: The architecture of Ghost module [11], which uses a series of cheap transformation operations to generate ghost feature maps, which results in significantly reduced computational complexity.

The model size of the L-Unet is only 8.7 M, which is four times less than that of the U-Net (35.5 M).

4. Experiments and Results

In this section, we conduct a series of experiments to evaluate the proposed LCC-Net's performance for semisupervised cardiac segmentation.

4.1. Dataset and Evaluation Measure. ACDC Dataset. We first utilize ACDC (Automated Cardiac Diagnosis Challenge) [12] dataset in our experiments, which belongs to a cardiac MR images segmentation challenge in MICCAI 2017. The ACDC dataset includes the short-axis cine-MRI of 150 subjects acquired from the University Hospital of Dijon using two MR scanners of different magnetic strengths. Left ventricle (LV), right ventricle (RV), and myocardium (MYO) were manually annotated by clinical experts on end-diastolic (ED) and end-systolic (ES) phase instants. The organizer of the ACDC splits the whole dataset into two subsets: (1) 100 subjects with available ground truth and (2) 50 subjects without ground truth for online testing.

We use the 100 labeled subjects (including 1902 image slices) for model evaluation. We randomly selected 20 subjects (containing 380 slices) as the testing set. The remaining 80 subjects are used as the union of the labeled data and unlabeled data. Specifically, we randomly select K subjects (2, 4, 6, and 10) for model training and the remaining 80- K subjects as the unlabeled data.

Evaluation Criteria. Our experiments utilize the Dice Coefficient (DICE) and Hausdorff distance (HD) as the evaluation criteria. Given the ground truth X and the prediction Y, DICE, which evaluates the region overlap of different segmentations, is defined as

$$\mathrm{DICE} = \frac{2 \bullet |X \cap Y|}{|X \cup Y|}. \tag{5}$$

The HD is defined as

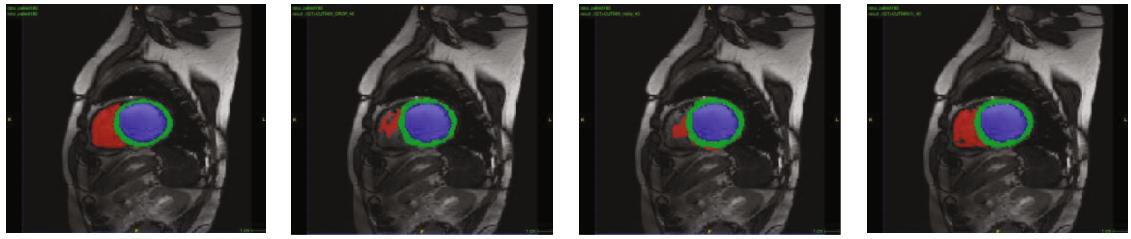
$$\mathrm{HD}(X, Y) = \max\left\{\max_{a \in X} E(a, Y), \max_{b \in Y} E(b, X)\right\}, \tag{6}$$

where $E(a, X)$ is the Euclid distance between a and X.

4.2. Implementation Details. We implemented our experiments on the framework of PyTorch [37] on one GTX 1080 GPU with 8 G memory. We used the adaptive moment estimation (Adam) optimizer with the learning rate of 5×10^{-4} initially, decreasing by 0.5 in epochs 200, 1000, 1500, 1800, and 2100. Moreover, the batch size was set as 4 because of the limitation of the GPU. The maximum epochs of iterations were set as 3000, and $\lambda_{\max}$ was set as 0.4. Data

Table 1: Comparative study of the proposed LCC-Net on the ACDC dataset. We randomly selected 2 subjects as the labeled data and the remaining 78 subjects as the unlabeled data. The models are tested on 20 unseen subjects. P_C denotes the perturbations in the feature space, P_N denotes the noise perturbation in the feature space, and P_D denotes the dropout perturbation in the feature space.

Method	Labeled data	DICE (%)				Hausdorff (mm)			
		LV	RV	MYO	Mean	LV	RV	MYO	Mean
U-Net (upbound)	80 subjects	93.2	85.8	88.9	89.3	2.2	4.8	2.8	3.3
U-Net (baseline)	2 subjects	76.1	24.7	69.1	56.6	9.2	17.0	11.4	12.5
L-Unet		69.9	36.7	60.2	55.6	10.7	19.2	12.7	14.2
LCC-Net w/o P_C		78.2	52.0	69.6	66.6	7.8	16.6	9.8	11.4
LCC-Net w/o P_N		78.0	54.8	70.7	67.8	5.8	11.3	7.8	8.3
LCC-Net w/o P_D		80.6	53.6	73.0	69.0	5.6	11.9	6.9	8.1
LCC-Net		82.0	58.1	73.0	71.0	4.3	13.9	6.3	8.2



(a) Ground truth (b) LCC-Net w/o P_N (c) LCC-Net w/o P_D (d) LCC-Net

Figure 5: Visualization of the performance of the ablated versions of our LCC-Net.

Table 2: The impact of the number of the labeled subjects. The results are tested on the ACDC dataset.

Labeled data	Method	LV (%)	RV (%)	MYO (%)	Mean (%)
2 subjects	U-Net (baseline)	76.1	24.7	69.1	56.6
	LCC-Net	82.0	58.1	73.0	71.0
4 subjects	U-Net (baseline)	79.6	51.7	71.1	67.5
	LCC-Net	85.0	69.3	76.9	77.1
6 subjects	U-Net (baseline)	83.4	62.3	74.0	73.2
	LCC-Net	85.0	74.5	79.3	79.6
10 subjects	U-Net (baseline)	82.1	70.1	76.6	76.3
	LCC-Net	87.4	77.0	82.8	82.4
Fully supervised (80 subjects)	U-Net (baseline)	93.2	85.8	88.9	89.3

augmentation, including affine transform, random rotation, and random intensity shift, was used. All the images were resized to 160×160, and the intensity range of each image was rescaled to [0, 1].

4.3. Segmentation Performance

4.3.1. Comparative Results of the LCC-Net. We first conduct a comparative study to identify the effectiveness of the critical components in the proposed model, including the backbone network L-Unet, the dropout decoder D_D, the noise decoder D_N, and the input space perturbation P_C. Specifically, we randomly select $K = 2$ subjects (40 slices) as the labeled data and the remaining 78 subjects as the unlabeled data, which are used for model training.

Table 1 summarizes the results of the comparative studies. The results of 7 network and data settings are reported: (1) the upbound, i.e., the U-Net trained with all the 80 labeled data; (2) the U-Net as the baseline, which is trained from scratch using the labeled data with standard data augmentations; (3) the L-Unet, which is also trained from scratch using the labeled data with standard data augmentations; (4) the LCC-Net w/o P_C, which is trained on both the labeled and unlabeled data without the input space perturbation P_C; (5) the LCC-Net w/o P_N, which is trained on both the labeled and unlabeled data without the noise decoder D_N; (6) the LCC-Net w/o P_D, which the LCC-Net without the dropout decoder D_D; (7) the full LCC-Net.

As shown in Table 1, when training with only two labeled subjects, the U-Net has a mean performance drop of 32.7% in DICE and 9.2 mm in Hausdorff than the U-Net trained with 80 subjects. Rather than using more labeled data, we exploit the unlabeled data, which is much easier to collect. As illustrated in Table 1, by exploiting unlabeled data, the LCC-

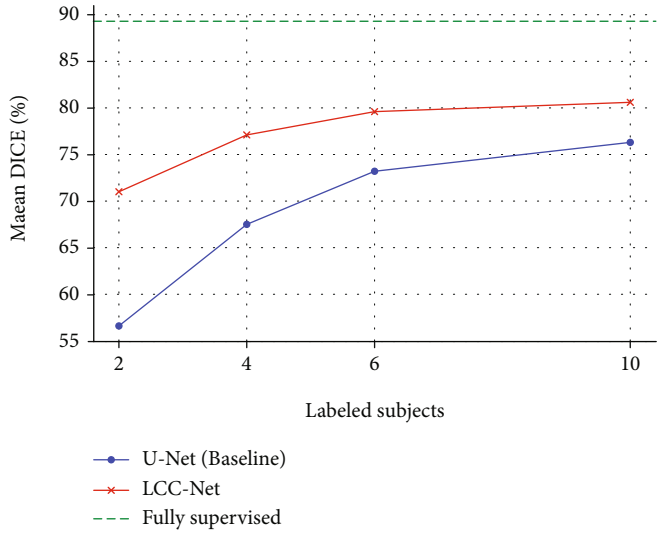


Figure 6: The impact of the number of the labeled subjects.

Table 3: The impact of the selection of the labeled subjects. The results are tested on ACDC dataset. Five samples are randomly selected, where each sample contains two labeled subjects as the labeled data for model training.

	DICE (%)							Hausdorff (mm)						
	(1)	(2)	(3)	(4)	(5)	Mean	Std	(1)	(2)	(3)	(4)	(5)	Mean	Std
LV	82.0	83.6	78.8	80.8	82.1	81.5	1.6	4.3	4.5	6.8	5.6	5.3	5.3	0.9
RV	58.1	64.0	51.0	52.9	54.6	56.1	4.6	13.9	9.1	12.4	11.9	10.1	11.5	1.7
MYO	73.0	74.1	73.6	73.9	75.5	74.0	0.8	6.3	5.8	6. 7	6.4	5.6	6.2	0.4
Mean	71.0	73.9	67.8	69.2	70.7	70.5	2.0	8.2	6.5	8.6	8.0	7.0	7.7	0. 8

Table 4: A comparison of the model complexity of the LCC-Net with different building blocks. The total inference time denotes the inference time on the whole testing set (402 images of 160×160).

LCC-Net with	Model size		FLOPs		Total
	Training	Testing	Training	Testing	Inference time
Standard convolution	81.5 M	35.5 M	329.8 G	102.8 G	20.8 s
Ghost module ($s = 4$)	25.2 M	8.7 M	91.0 G	26.2 G	17.3 s

Net outperforms the U-Net (baseline) by a large margin, i.e., 14.4% in the mean DICE and 4.3 mm in the mean Hausdorff over the three regions. LCC-Net without using the noise perturbation P_N and noise decoder D_N obtains a performance gain of 12.2% in DICE and 5.9 mm in Hausdorff over the L-Unet; LCC-Net without using the dropout perturbation P_D and dropout decoder D_D obtains a performance gain of 13.4% in DICE and 6.1 mm in Hausdorff over the L-Unet. Compared to the full LCC-Net, the LCC-Net without using the input space perturbation P_C shows a mean performance drop of 4.4% in DICE and 3.2 mm in Hausdorff, which indicates the effectiveness of the input space perturbation P_C. However, with only two labeled subjects for model training, the semisupervised model's performance is still significantly lower than the fully supervised U-Net. Figure 5 provides a visual comparison of the LCC-Net without P_N, LCC-Net without P_D, and our LCC-Net. Visually, the LCC-Net shows significantly better results than the other two methods.

4.3.2. The Impact of the Number of the Labeled Subjects. Since our method is a semisupervised method, it is crucial to identify the impact of the size of the labeled training dataset. To this end, we trained our model with different choices of K, i.e., 2, 4, 6, and 10 subjects. Table 2 summarizes the experimental results. The results with U-Net under different settings, including the fully supervised setting (80 labeled subjects), are also reported. As can be expected, with increasingly more labeled data for model training, the performance becomes much higher. With the different choices of K, our semisupervised model consistently outperforms the U-Net. Remarkably, using only four labeled subjects, our model outperforms the U-Net trained on ten labeled subjects. Using ten labeled subjects for training, the LCC-Net achieves a mean performance of 82.4%, which is 6.1% higher in mean DICE than the U-Net. Figure 6 demonstrates a further comparison of the proposed model and the U-Net, which shows the effectiveness of our model.

4.3.3. The Impact of the Selection of the Labeled Subjects. To identify the robustness of the proposed model over the different selections of the label data. To this end, we randomly selected five samples and calculated the mean performance and the standard variance. Here, each sample contains two subjects as the labeled data. The results are reported in Table 3. Although each sample size is very small (2 subjects), our model shows relatively stable performance.

4.4. Model Complexity. Model complexity is typically measured by the number of trainable network parameters (i.e., model size) and the floating-point operations (FLOPs). The model complexity of our model is summarized in Table 4. Our model obtained significantly reduced model size and FLOPs at both the training stage and testing stage by replacing standard convolutions with the lightweight module. Therefore, our model requires less computation cost for each training step and inference step, resulting in higher computational efficiency. The inference time at the testing stage is a critical measure in practical usage. As shown in Table 4, with reduced FLOPs, the proposed LCC-Net involves a shorter inference time than the LCC-Net using standard convolutions.

5. Conclusion

In this paper, we presented a lightweight cross-consistent network for semisupervised cardiac MR image segmentation. We leveraged the unlabeled data during supervised segmentation model learning and encourage segmentation consistency on all data under different perturbations with two unsupervised consistency losses. To achieve a lightweight model, we replaced the standard convolutions with a lightweight module. Extensive comparison experiments with a public dataset demonstrated that our architecture achieved promising performance with only two labeled subjects.

Despite the improved results, there are still more applicable perturbations in semisupervised segmentation. Thus, exploring more efficient perturbations is a significant work in the future.

Data Availability

The data used in our experiments are available at https://www.creatis.insa-lyon.fr/Challenge/acdc/.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the NSFC under Grant 11771160, the Fund of HQU (ZQN-PY411), and by STPF under Grant 2019H0016.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Springer, 2015.

[3] C. Chen, C. Qin, H. Qiu et al., "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, 2020.

[4] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning*, vol. 3, no. 2, 2013ICML, 2013.

[5] Y. Grandvalet and Y. Bengio, *Semi-Supervised Learning by Entropy Minimization*, Advances in Neural Information Processing Systems, 2005.

[6] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2020.

[7] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," in *IEEE Transactions on Neural Networks*, vol. 20no. 3, p. 542, 2009.

[8] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, *Semi-Supervised Learning with Ladder Networks*, Advances in Neural Information Processing Systems, 2015.

[9] A. Tarvainen and H. Valpola, *Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results*, Advances in Neural Information Processing Systems, 2017.

[10] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, https://arxiv.org/abs/1610.02242.

[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: more features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

[12] O. Bernard, A. Lalande, C. Zotti et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[13] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, "Cardiac MRI segmentation with strong anatomical guarantees," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 632–640, Springer, 2019.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations*, 2014.

[15] M. Khened, V. Alex, and G. Krishnamurthi, "Densely Connected Fully Convolutional Network for Short-Axis Cardiac Cine MR Image Segmentation and Heart Diagnosis Using Random Forest," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 140–151, Springer, 2017.

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densenet: densely connected convolutional networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 30, pp. 82–84, 2017.

[17] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-Balanced Deep Neural Network for Automatic Ventricular

Structure Segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 152–160, Springer, 2017.

[18] G. Simantiris and G. Tziritas, "Cardiac mri segmentation with a dilated cnn incorporating domain-specific constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1235–1243, 2020.

[19] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic Cardiac Disease Assessment on Cine-MRI via Time-Series Segmentation and Domain Specific Features," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 120–129, Springer, 2017.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2016.

[21] B. Liu, Z. Wu, H. Hu, and S. Lin, "Deep metric transfer for label propagation with limited annotated data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[22] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613, Springer, 2019.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[24] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, https://arxiv.org/abs/1802.07934.

[25] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 370–378, Springer, 2018.

[26] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.

[27] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818, Springer, 2019.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: beyond empirical risk minimization," 2017, https://arxiv.org/abs/1710.09412.

[29] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[31] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Consistency regularization and cutmix for semi-supervised semantic segmentation," 2019, https://arxiv.org/abs/1906.01916.

[32] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[34] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4373–4382, 2017.

[35] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[36] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

[37] A. Paszke, S. Gross, S. Chintala et al., "Automatic differentiation in pytorch," in *31st Conference on Neural Information Processing Systems (NIPS 2017)- Workshop*, 2017.