Contrast Agent-free Synthesis and Segmentation of Ischemic Heart Disease Images using Progressive Sequential Causal GANs

Chenchu Xu[a,*], Lei Xu[b,*], Pavlo Ohorodnyk[a], Mike Roth[a], Bo Chen[c,**], Shuo Li[a,**]

[a] *Department of Medical Imaging, Western University, London ON, Canada*
[b] *Department of Radiology, Beijing AnZhen Hospital, Beijing, China*
[c] *School of Health Science, Western University, London ON, Canada*

Abstract

The elimination of gadolinium contrast agent (CA) injections and manual segmentation are crucial for ischemic heart disease (IHD) diagnosis and treatment. In the clinic, CA-based late gadolinium enhancement (LGE) imaging and manual segmentation remain subject to concerns about potential toxicity, interobserver variability, and ineffectiveness. In this study, progressive sequential causal GANs (PSCGAN) are proposed. This is the first one-stop CA-free IHD technology that can simultaneously synthesize an LGE-equivalent image and segment diagnosis-related tissues (i.e., scars, healthy myocardium, blood pools, and other pixels) from cine MR images. To this end, the PSCGAN offer three unique properties: 1) a progressive framework that cascades three phases (i.e., priori generation, conditional synthesis, and enhanced segmentation) for divide-and-conquer training synthesis and segmentation of images. Importantly, this framework leverages the output of the previous phase as a priori condition to input the next phase and guides its training for enhancing performance, 2) a sequential causal learning network (SCLN) that creates a multi-scale, two-stream pathway and a multi-attention weighing unit to extract spatial and temporal dependencies from cine MR images and effectively select task-specific dependence. It also integrates the GAN architecture to leverage adversarial training to further facilitate the learning of interest dependencies of the latent space of cine MR images in all phases; and 3) two

[*] These authors contributed equally to this work.
[**] Corresponding author
Email address: slishuo@gmail.com (Shuo Li)

specifically designed self-learning loss terms: a synthetic regularization loss term leverages the spare regularization to avoid noise during synthesis, and a segmentation auxiliary loss term leverages the number of pixels for each tissue to compensate for discrimination during segmentation. Thus, the PSCGAN gain unprecedented performance while stably training in both synthesis and segmentation. By training and testing a total of 280 clinical subjects, our PSCGAN yield a synthetic normalization root-mean-squared-error of 0.14 and an overall segmentation accuracy of 97.17%. It also produces a 0.96 correlation coefficient for the scar ratio in a real diagnostic metric evaluation. These results proved that our method is able to offer significant assistance in the standardized assessment of cardiac disease.

Keywords: Gadolinium contrast agents, Synthesis, Sequential learning, Ischemic heart disease, Progressive framework

1. Introduction

1.1. Clinical concerns about contrast-agents and manual segmentation

Gadolinium-based contrast agents (CA) imaging and manual segmentation of diagnosis-related tissues are essential parts of the current ischemic heart disease (IHD) treatment workflow in cardiac radiology (Beckett et al., 2015; Bijnens et al., 2007). CA imaging uses chemical substances in MR scans (Moon et al., 2004). After the CA is injected into the body, CA imaging produces a late gadolinium enhancement (LGE) image to illustrate IHD scars that are invisible under regular MR imaging and improves the clarity of other internal and surrounding cardiac tissues (i.e., muscles, cavities, and even blood). Furthermore, manual segmentation delineates diagnosis-related tissues (scars, myocardium, etc.). After the CA imaging, manual segmentation helps radiologists to segment multiple cardiac tissues, and the subsequent quantitative evaluation of these segmented tissues results in various diagnosis metrics to accurately report the presence of the progression of IHD (Fox et al., 2010).

However, with this workflow (i.e., CA imaging first followed by manual segmentation), there are still faces concerns regarding toxicity, high interobserver variability, and ineffectiveness (Kali et al., 2014). 1) CAs have been highlighted in numerous clinical papers showing their potential toxicity, retention in the human body, and importantly, their potential to induce fatal nephrogenic systemic fibrosis (Ordovas and Higgins, 2011). 2) Manual segmentation has well-known issues regarding high interobserver variability and non-reproducibility, which are caused by the difference in expertise among clinicians (Ordovas and Higgins, 2011). 3)

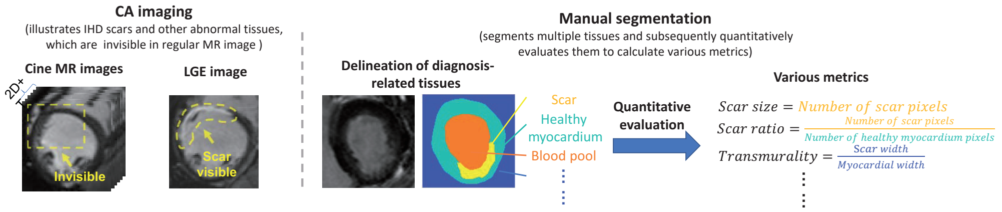
CA imaging
(illustrates IHD scars and other abnormal tissues, which are invisible in regular MR image)
Cine MR images
2D+T
Invisible
LGE image
Scar visible
Manual segmentation
(segments multiple tissues and subsequently quantitatively evaluates them to calculate various metrics)
Delineation of diagnosis-related tissues
Scar
Healthy myocardium
Blood pool
Quantitative evaluation
Various metrics
$Scar\ size = Number\ of\ scar\ pixels$
$Scar\ ratio = \frac{Number\ of\ scar\ pixels}{Number\ of\ healthy\ myocardium\ pixels}$
$Transmurality = \frac{Scar\ width}{Myocardial\ width}$

Figure 1: Gadolinium-based contrast agent (CA) imaging and scar manual segmentation are essential parts of the current ischemic heart disease (IHD) treatment workflow.

CA imaging followed by segmentation leads to additional time and effort for both patients and clinicians, as well as high clinical resource costs (labor and equipment). (Ingkanisorn et al., 2004).

1.2. *Clinical limitations of existing initial CA-free scar segmentation methods*

To date, a few initial CA-free and automatic segmentation methods have been reported(Suinesiaputra et al., 2017; Wong et al., 2016). However, even the state-of-the-art methods only produce a binary scar image that fails to provide a credible diagnosis (Xu et al., 2018a,b). As shown in Figures 2 and 3, this binary scar image can only indicate two categories of pixels: scar and background. This limited resolution thus fails to highlight all the essential tissues (e.g., myocardium and healthy myocardium, blood pool) recommended according to the clinical protocols of comprehensive IHD evaluation. Subsequently, it fails to help radiologists quantitatively assess multiple tissues to obtain the most powerful metrics for a credible IHD diagnosis (e.g., scar ratio = size of the scar/size of the myocardium). Because the use of multiple metrics based on multiple tissues results in far greater accuracy than using only a metric based on scar tissue alone in a credible IHD diagnosis (Zhang et al., 2019), the limitations of existing segmentation methods need to be addressed.

Thus, clinicians urgently desire the development of more advanced CA-free technology that should simultaneously produce an LGE-equivalent image (i.e., an image that is equivalent to an LGE image in terms of usefulness in an IHD diagnosis or from which clinical metrics can be obtained without CA injections) and a segmented image (including all diagnosis-related tissues, i.e., scar, healthy myocardium, and blood pools, as well as other pixels) (Leiner, 2019).

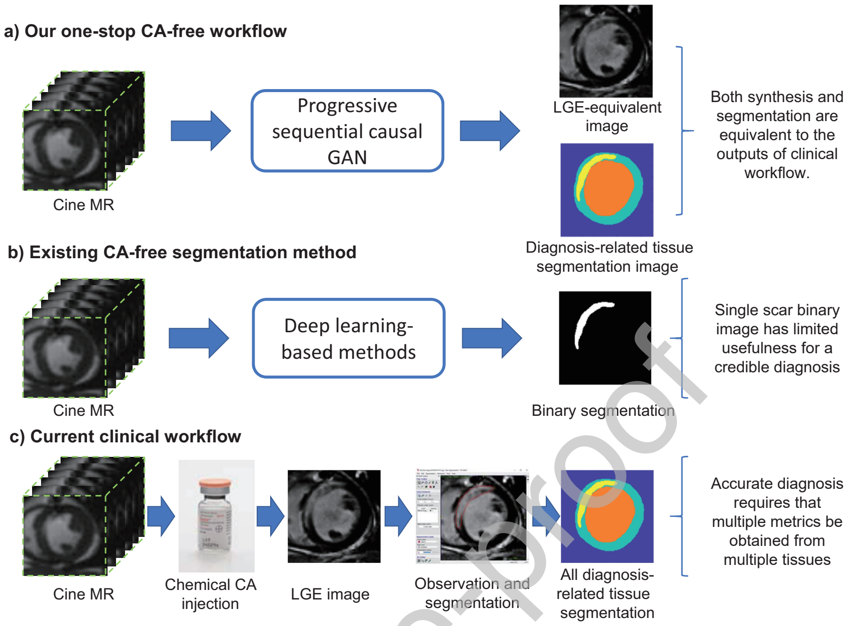
a) Our one-stop CA-free workflow
Cine MR
Progressive sequential causal GAN
LGE-equivalent image
Diagnosis-related tissue segmentation image
Both synthesis and segmentation are equivalent to the outputs of clinical workflow.
b) Existing CA-free segmentation method
Cine MR
Deep learning-based methods
Binary segmentation
Single scar binary image has limited usefulness for a credible diagnosis
c) Current clinical workflow
Cine MR
Chemical CA injection
LGE image
Observation and segmentation
All diagnosis-related tissue segmentation
Accurate diagnosis requires that multiple metrics be obtained from multiple tissues

Figure 2: PSCGAN as a one-stop CA-free technology for the simultaneous synthesis of LGE-equivalent images and the segmentation of all diagnosis-related tissues (including scar, healthy myocardium and blood pools, as well as other pixels) for IHD diagnosis. It provides an accurate IHD diagnostic output that is equivalent to clinical CA-based imaging and manual segmentation by experts, rather than only a binary scar image as produced by existing state-of-the-art methods.

1.3. Technical challenges of LGE equivalent image synthesis and multiple diagnosis-related tissue segmentation

However, it is very challenging to synthesize an LGE-equivalent image and accurately segment all the diagnosis-related tissues (i.e., scar, healthy myocardium and blood pools) from 2D+T cine MR images. 1) The pixel-level understanding of LGE images by representation learning of the 2D+T cine MR images faces the issue of numerous instances. The differences in the enhancement effects of the CAs on different cardiac cells result in each of the numerous pixels of the LGE image requiring a definite non-linear mapping from the cine MR images. 2) Representation learning of the 2D+T cine MR has a number of high-complexity issues. The time series characteristics of 2D+T cine MR images result in each non-linear mapping requiring a complex mixing of the spatial and temporal dependencies of a mass of pixels in the images, especially since these pixels often have high local variations (Luc et al., 2016). 3) More importantly, a pixel-level understanding of LGE images is needed to differentiate between pixels that have very similar appearances(Xu et al., 2017). The highly similar intensity of pixels within the tissue on an LGE image often results in high similarities between the learned spatial and temporal dependencies of these pixels and often causes interference and inaccuracy during mixing. The combination of all three issues makes the synthesis and segmentation of LGE-equivalent images incredibly challenging.

1.4. Existing progressive networks

Recently, progressive generative adversarial networks (GAN) have shown great potential in the tasks of image synthesis and segmentation (Huang et al., 2017; Karras et al., 2017; Zhang et al., 2018b). Progressive GAN inherit the advantage of adversarial semi-supervised learning from GAN to effectively learn to map from a latent space to a data distribution of interest. More importantly, the progressive framework of such progressive GAN stacks multiple sub-GAN networks as different phases to take advantage of the result of the previous phase to guide the performance of the next phase and greatly stabilize training. However, current progressive GAN are designed to train on a single task because they lack a two-task generation scheme to simultaneously handle the synthesis task and segmentation task.

1.5. Progressive sequential causal GANs

In this paper, we propose a progressive sequential causal GAN (PSCGAN) as a one-stop CA-free technology that can simultaneously synthesize an LGE equivalent image and segment a diagnosis-related tissue segmentation image from cine MR images to diagnose IHD. To the best of our knowledge, this is the first technology to synthesize an image equivalent to a CA-based LGE-image and to segment multiple tissues equivalently to the manual segmentation performed by experts, as well as offer simultaneous synthesis and segmentation.

Our PSCGAN innovatively build three phases in a step-by-step cascade of three independent GANs (i.e., the priori generation GAN, the conditional synthesis GAN, and the enhanced segmentation GAN). The first phase uses the priori generation GAN to train the network on a coarse tissue mask; the second phase uses the conditional synthesis GAN to synthesize the LGE-equivalent image; and the third phase uses the enhanced segmentation GAN to segment the diagnosis-related tissue image. Importantly, the PSCGAN create a pipeline to leverage the commonalities between the synthesis task and the segmentation task. This pipeline takes the pixel categories and distributions in the coarse tissues mask as a priori condition to guide the LGE-equivalent image synthesis. It also takes the fine texture in the LGE-equivalent image as a priori condition to guide the diagnosis-related tissue segmentation. PSCGAN use these two reciprocal guidances between the two tasks to gain an unprecedentedly high performance in both tasks while performing stable training.

Our PSCGAN further implement the following novelties: 1) a novel sequential causal learning network (SCLN). The SCLN creatively builds a two-stream dependency-extraction pathway and a multi-attention weighing unit. The two-stream pathway multi-scale extracts the spatial and temporal dependencies separately in the spatiotemporal representation of images to include the short-range to the long-range scale variants; the multi-attention weighing unit computes the responses within and between spatial and temporal dependencies at the task output as a weight and mixes them according to the assigned weights. This network also integrates with GAN architecture to further facilitate the learning of interest dependencies of the latent space of cine MR images in all phases, and 2) the adoption of two specially designed loss terms, i.e., a synthetic regularization loss term and a self-supervised segmentation auxiliary loss term for optimizing the synthesis task and the segmentation task respectively. The synthetic regularization loss term uses a spare regularization learned from the group relationship between the intensity of the pixels to avoid the noise during the synthesis, thereby improving the quality of the synthesized image, while the self-supervised segmentation auxiliary loss term uses the number of pixels in each tissue to balance the output rather than only the shape of the tissues to improve the discrimination performance of the segmented image and thereby improve the segmentation accuracy.

1.6. Contribution

In summary, the main contributions of this work are as follows:

- For the first time, a CA-free synthesis and segmentation method is proposed. This method eliminates the CA-associated health risks and streamlines the clinical workflows.

- A novel sequential causal learning framework is proposed. This framework strengthens the spatiotemporal repesentation learning of time-series images by gaining task-specific spatiotemporal dependencies.

- A progressive framework cascading three reciprocal GANs is proposed for both image synthesis and segmentation. It exploits the commonalities of the synthesis task and the segmentation task, as well as obtaining high performance and stable training.

2. Related Work

2.1. Existing IHD methods for CA injection and manual segmentation

Currently, there is no method for both synthesizing LGE-equivalent images and segmenting all diagnosis-related tissues directly from cine MR images. Early, traditional CA-free IHD-diagnosing methods, such as energy-based and statistical shape model-based methods (Ledesma-Carbayo et al., 2005; Suinesiaputra et al., 2017), cannot perform automatic segmentation. These methods only produce image-level IHD classification or region-level IHD scar localization from cine MR images; therefore, radiologists often need to further manually segment these classification or localization results for diagnosis. With the introduction of deep learning, some CA-free IHD-diagnosing methods, such as 3DConv-based or LSTM-based methods, have been used to segment a pixel-level scar from cine MR images and have been reported by the radiology community in a real clinical setting (as mentioned in section 1.2) (Duchateau et al., 2016; Xu et al., 2018b; Zhang et al., 2019; Tan et al., 2012).

Moreover, existing IHD-diagnosing methods, even the state-of-the-art one proposed by us (Xu et al., 2018a), are inefficient in the representation learning of cine MR images. 1) Existing methods still must contend with a fixed local observation in both spatial dependency and temporal dependency extraction (e.g., only adjacent temporal frames of optical flow and a fixed spatial convolutional kernel size for deep learning). However, pixels in 2D+T cine MR images often have high local variations (i.e., different positions and motion ranges in different regions and timestamps) (Luc et al., 2016; Su et al., 2020). 2) Current spatial-temporal feature learning methods still struggle with constant learning weights during the mixing of spatial dependencies with temporal dependencies (e.g., both 3DConv and ConvLSTM often simply treat the two dependencies on each pixel as equal during learning) (Xu et al., 2017). However, different pixels have different selection requirements in terms of temporal dependencies and spatial dependencies (Tan et al., 2013b,a).

2.2. Generative adversarial networks

GANs (Goodfellow et al., 2014) have become one of the most promising deep learning architectures for either image segmentation tasks or synthesis tasks in recent years. However, GANs may produce inefficient and unstable results when two or more tasks need to be solved at the same time. GAN comprises two networks, a generator and a discriminator, where one is pitted against the other. The generator network learns to map from a latent space to a data distribution of interest, while the discriminator network distinguishes the candidates produced by the generator from the true data distribution. However, a GAN may learn an erroneous data distribution or a gradient explosion when the latent space of the distributions of two tasks interfere with each other. Conditional GAN, a type of GAN implementation, has the potential to learn reciprocal commonalities of the two tasks to avoid interferes with each other because of its considerable flexibility in how two hidden representations are composed (Mirza and Osindero, 2014; Isola et al., 2017). In conditional GAN, a conditioned parameter y is added to the generator to generate the corresponding data using the following equation:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_{z}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))] \tag{1}$$

where $p_{\text{data}}(\boldsymbol{x})$ represents the distribution of the real data and $p_{\boldsymbol{z}}$ represents the distribution of the generator.

2.3. Attention model

The attention model successfully weighs the positions that are highly related to the task (Bahdanau et al., 2014), thereby improving the performance of the application in various tasks (Zhou et al., 2016; Vaswani et al., 2017). It is inspired from the way humans observe images, wherein more attention is paid to a key part of the image in addition to understanding an image as a whole. Such a model uses convolutional neural networks as basic building blocks and calculates long-range representations that respond to all positions in the input and output images. It then determines the key parts that have high responses in the long-range representations and weights these parts to motivate the networks to better learn the images. In particular, recent work on attention models embedded an auto regressive model to achieve image synthesis and segmentation by calculating the response at a position in a sequence through attention to all positions within the same sequence (Zhang et al., 2018a). This model has also been integrated into GANs by attending to internal model states to efficiently find global, long-range dependencies within the internal representations of the images. Importantly, the attention model has been formalized as a non-local operation to model the spatial-temporal dependencies in video sequences (Wang et al., 2018). Despite this progress, the attention model has not yet been explored for the internal effects of different spatial and temporal combinations on synthesis and segmentation in the context of GANs.

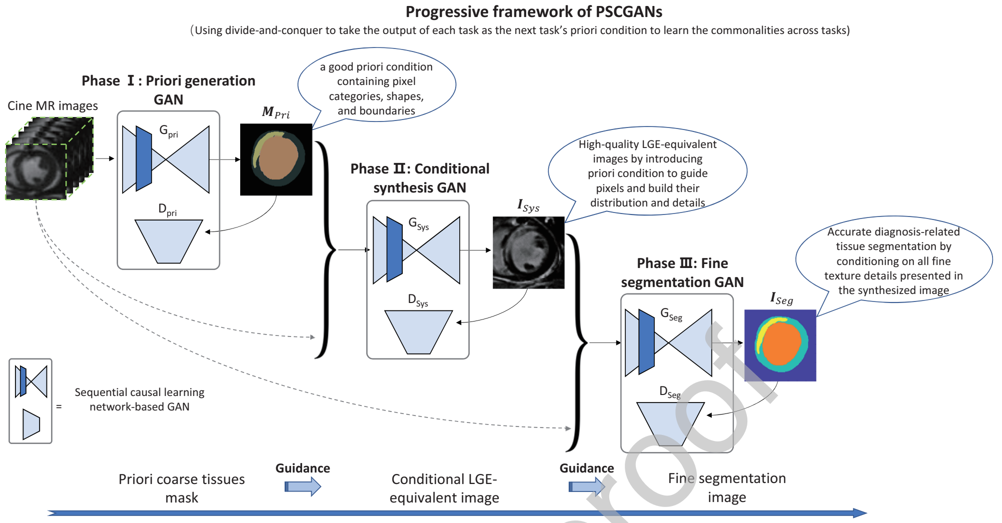
Progressive framework of PSCGANs
（Using divide-and-conquer to take the output of each task as the next task's priori condition to learn the commonalities across tasks)
Phase Ⅰ: Priori generation GAN
Cine MR images
a good priori condition containing pixel categories, shapes, and boundaries
M_{Pri}
G_{pri}
D_{pri}
Phase Ⅱ: Conditional synthesis GAN
High-quality LGE-equivalent images by introducing priori condition to guide pixels and build their distribution and details
I_{Sys}
G_{Sys}
D_{Sys}
Phase Ⅲ: Fine segmentation GAN
Accurate diagnosis-related tissue segmentation by conditioning on all fine texture details presented in the synthesized image
I_{Seg}
G_{Seg}
D_{Seg}
= Sequential causal learning network-based GAN
Priori coarse tissues mask
Guidance
Conditional LGE-equivalent image
Guidance
Fine segmentation image

Figure 3: The SCLN creatively builds a two-stream pathway (i.e., a spatial perceptual pathway and a temporal perceptual pathway) to separately extract multi-scale and multi-level spatial and temporal dependencies from cine MR images. Then, it also builds a multi-attention weighing unit to compute and select the task-specific dependencies within and between these two dependencies.

3. Overview of PSCGAN

As depicted in Figure 3, PSCGAN cascade three GANs to build three phases and connect them by taking the output of the previous GAN as an input of the next GAN. Moreover, to reduce the randomness during training, all three GANs encode the cine MR images by using the same foundational network architecture, a SCLN-based GAN (Sect. 4.2) that includes an encoder-decoder generator and a discriminator to specially design and handle time-series images. Thus, PSCGAN not only have great training stability by using divide-and-conquer to separate the segmentation task and synthesis task into different phases but also undergo effective training by progressively taking the output of the previous phase as the priori condition input to guide the next phase .

Phase I: priori generation GAN (Sect.5.1). This phase uses the priori generation GAN (Pri) to generate a coarse tissue mask $\mathbf{M_{Pri}}$ from the cine MR images $\mathbf{X}$ by adversarial training. This coarse segmented image is a rich priori condition, as it contains all pixel categories and tissue shapes, locations, and boundaries.

Phase II: conditional synthesis GAN (Sect.5.2). This phase uses the conditional synthesis GAN (Sys) to integrate the coarse tissue mask and the cine MR image to build a conditional joint mapping to use the obtained pixel attributes and distributions from the mask to guide the image synthesis to generate a high-quality LGE-equivalent image $\mathbf{I_{sys}}$.

Phase III: enhanced segmentation GAN (Sect.5.3). This phase uses the enhanced segmentation GAN (Seg) to introduce the synthesized image from Sys as a priori condition to generate the diagnosis-related tissue segmentation image I_{Seg}. The synthesized image and all detailed textures effectively guide the classification of the tissue boundary pixels.

4. Sequential causal learning network (SCLN)-based GAN

The core of the SCLN-based GAN is our newly proposed SCLN. An SCLN is a novel spatiotemporal representation learning framework for the 2D+T time-series image. It has the ability to select the task-specific dependence between and within the extracted spatial and temporal dependencies from the 2D+T time-series image. Thus, in our work, the SCLN improves the spatiotemporal representation learning of 2D+T cine MR images and facilitates the accuracy of the pixel-level nonlinear mapping from the 2D T cine MR images to synthesis and segmentation. Moreover, by integrating an SCLN into the GAN architecture as the encoder of the cine MR images in the generator, the SCLN-based GAN improves the learning effectiveness of the interest distribution from the latent space of the cine MR images, thereby effectively improving the generating performance on adversarial training.

4.1. Sequential causal learning network (SCLN)

The SCLN consists of a two-stream structure that includes a spatial perceptual pathway and a temporal perceptual pathway and a multi-attention weighing unit. The SCLN leverages the two-stream structure to flexibly divide the spatial dependence and the temporal dependence in the 2D+T time-series image into two independent learning pathways. It enables both the spatial dependence and temporal dependence learning to be focused by their corresponding pathway, thereby avoiding the interference between these two types of dependencies during learning. Moreover, the SCLN leverages the multi-attention weighing unit to weigh both the spatial dependence and temporal dependence, and it performs feature selection. It produces task-specific dependencies through the flexible mixing of the spatial dependence and temporal dependence by learnable weights, rather than mixing them based on constant learning weights, as in current spatiotemporal

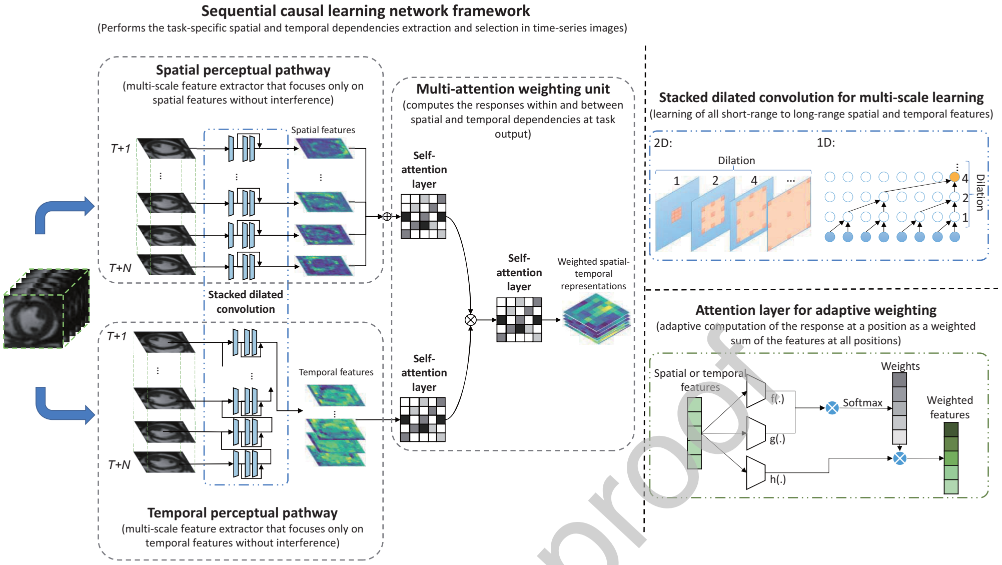
Sequential causal learning network framework
(Performs the task-specific spatial and temporal dependencies extraction and selection in time-series images)
Spatial perceptual pathway
(multi-scale feature extractor that focuses only on spatial features without interference)
T+1
T+N
Spatial features
Stacked dilated convolution
T+1
T+N
Temporal features
Temporal perceptual pathway
(multi-scale feature extractor that focuses only on temporal features without interference)
Multi-attention weighting unit
(computes the responses within and between spatial and temporal dependencies at task output)
Self-attention layer
Self-attention layer
Self-attention layer
Weighted spatial-temporal representations
Stacked dilated convolution for multi-scale learning
(learning of all short-range to long-range spatial and temporal features)
2D:
Dilation
1
2
4
...
1D:
4
2
1
Dilation
Attention layer for adaptive weighting
(adaptive computation of the response at a position as a weighted sum of the features at all positions)
Spatial or temporal features
f(.)
g(.)
h(.)
Softmax
Weights
Weighted features

Figure 4: SCLN creatively builds two-stream pathways (i.e. a spatial perceptual pathway and a temporal perceptual pathway) to separately extract multi-scale and multi-level spatial and temporal dependencies from cine MR images. It also builds the multi-attention weighing unit to respectively compute and select the task-specific dependence within and between these two dependencies.

learning methods. Thus, the SCLN strengthens the accuracy of the spatiotemporal dependencies, thereby improving the representation of the 2D+T time-series images.

4.1.1. *Two-stream structure for multi-scale spatial and temporal dependency extraction*

As shown in Figure 4, the spatial perceptual pathway and the temporal perceptual pathway use two independent, stacked dilated convolution (Yu and Koltun, 2015) as multi-scale extractors to focus the spatial dependencies and the temporal dependencies in the time-series images, respectively. Dilated convolution consists of sparse filters that use skip points during convolution to exponentially grow the receptive field to aggregate multi-scale context information. It improves the diversity of both spatial dependencies and temporal dependencies to include all the short-range to long-range scale variants. The 1D/2D dilated convolutions are formulated as follows:

$$\mathbf{1D} : (kernel *_l x)_t = \sum_{s=-\infty}^{\infty} kernel_s \cdot f_{t-ls} \tag{2}$$

$$\mathbf{2D} : (x *_l kernel)(p) = \sum_{s+lt=p} x(s) kernel(t) \tag{3}$$

where x is the 1D/2D signal/image, and l is the dilation rater.

In our work, the spatial perceptual pathway uses 2D dilated convolution (Yu and Koltun, 2015), and the temporal perceptual pathway uses 1D dilated convolution (Oord et al., 2016). The inputs of both pathways are cine MR images. The spatial perceptual pathway regards $2D + T$ cine MR images as multiple (time t to time $t + n$) independent 2D images. Each input image is learned by a 2D dilated convolution, where the number of 2D dilated convolution is the same as the number of frames. The output of the 2D dilated convolution in time t is the spatial feature convolved with the frame of time t only. Thus, the spatial feature of $2D + T$ cine MR images can be effectively captured when combining all 2D dilated convolution from time t to time $t + n$. By contrast, the spatial perceptual pathway regards $2D + T$ cine MR images as a whole 1D data. This 1D data is learned by 1D dilated convolutions according to its order, where the hidden units of the 1D dilated convolution that are the same length as the 1D form of each frame (the length of a 64x64 frame is 4096). The output of each 1D dilated convolution time t is the temporal feature convolved with the frame of time t and the earlier time in the previous layer. Thus, the temporal feature of $2D + T$ cine MR can be effectively captured when the 1D dilated convolution process reaches the time $t + n$.

Concretely, both pathways initially stack 6 dilated convolutions, and the corresponding dilation rate is [1, 1, 2, 4, 6, 8]. This setting allows the learned representation to include all 3×3 to 65×65 motion and deformation scales. Note that the stack number still varies with the spatial and temporal resolution of the time-series image during encoding. Moreover, both spatial and temporal perceptual pathways stack 3 stacked dilated convolutions (1D/2D) again to build a residual block framework for deepening the network layers and enriching hierarchical features (He et al., 2016). Both paths also adopt a causal padding to ensure that the output at time t is only based on the convolution operation at the previous time (Oord et al., 2016). This causal-based convolution means that there is no information leakage from the future to the past.

In summary, the **advantages** of this two-stream structure are as follows: 1) two pathways are used to focus on two aspect dependencies independently; 2) dilated convolution with residual blocks and shortcut connections are used to extract multiscale and multilevel dependencies and 3) causal padding is used to understand the time order within the dependencies.

4.1.2. Multi-attention weighing unit for task-specific dependence selection

The multi-attention weighing unit consists of three independent self-attention layers and an add operator to adaptively weigh the high-contribution dependences between and within spatial and temporal dependencies at the output to perform accurate task-specific dependence selection (Vaswani et al., 2017). Two self-attention layers first embed behind both the spatial perceptual pathway and the temporal perceptual pathway to adaptively compute the response of each pathways dependence at the output as their weights; then, the add operator element-wise fuses the weighed spatial and temporal dependencies; finally, the third self-attention layer determines which of the fused spatial-temporal dependences is the task-specific dependence. Concretely, the spatial dependencies from the spatial perceptual pathway are defined as $\mathcal{F}_{S_{Conv}} \in R^{C\times N}$, where C is the number of channels and N is the number of dependencies. The spatial self-attention layer first maps these spatial dependencies into two feature spaces $f(.) = W_f\mathcal{F}_{S_{Conv}}$ and $g(.) = W_g\mathcal{F}_{S_{Conv}}$. It calculates the weight α_i to the i_{th} dependencies, where $\alpha = (\alpha_{\mathbf{1}}, \alpha_2, \ldots, \alpha_j, \ldots, \alpha_N) \in R^{C\times N}$:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{i=1}^{N}\exp(s_i)}, \text{ where } s_i = f\left(\mathcal{F}_{S_{Conv}i}\right)^T g\left(\mathcal{F}_{S_{Conv}i}\right) \tag{4}$$

The weighed spatial dependencies $\alpha\mathcal{F}_{S_{Conv}}$ are as follows:

$$v\left(\sum_{i=1}^{N}\alpha_i h\left(\mathcal{F}_{S_{Conv}i}\right)\right), \tag{5}$$

$$h\left(\mathcal{F}_{S_{Conv}i}\right) = \boldsymbol{W}_h\mathcal{F}_{S_{Conv}i}, v\left(\mathcal{F}_{S_{Conv}i}\right) = W_v\mathcal{F}_{S_{Conv}i} \tag{6}$$

where W_g, W_f, W_h, W_v are the learned weight matrices. For memory efficiency, $\{W_g, W_f, W_h, W_v\} \in \mathbb{R}^{\widetilde{C}\times C}$, where $\widetilde{C}$ is the reduced channel number and $\widetilde{C} = C/8$. Note that 8 is a hyperparameter.

By the same token, the temporal self-attention layer enhances the temporal dependencies $\mathcal{F}_{T_{Conv}}$ from the temporal perceptual path to an attention-weighted $\beta\mathcal{F}_{T_{Conv}} \in R^{C\times N}$, where $\beta = (\beta_{\mathbf{1}}, \beta_2, \ldots, \beta_j, \ldots, \beta_N) \in R^{C\times N}$ are the weights of the temporal dependencies.

The add operator elementwise fuses the weighed spatial dependencies and temporal dependencies:

$$\mathcal{F}_{ST_{Conv}} = \alpha \mathcal{F}_{S_{Conv}} + \beta \mathcal{F}_{T_{Conv}} \tag{7}$$

The fused self-attention layer weighs the fused spatial-temporal dependencies $\mathcal{F}_{ST_{Conv}}$. The output of this layer is $O_{ST_{Conv}} \in R^{C\times N}$. This output further adds the input of the map layer after modification with a learnable scalar γ. Therefore, the final output is given by $\gamma O_{St_{Conv}} + \mathcal{F}_{ST_{Conv}}$.

4.2. Implementation of an SCLN-based GAN for the basic network architecture at all phases

This network stacks 4 SCLNs and 4 corresponding up-sampling blocks to build a generator. The network further stacks 5 convolutional layers to build a discriminator. Both the generator and discriminator use conditional adversarial training to effectively perform the segmentation and synthesis.

As shown in Figure 5, the generator is an encode-decode 2D+T to 2D framework modified from U-Net (Ronneberger et al., 2015). It first encodes the input $X \in R^{25\times 64\times 64\times 1}$ (25 frames, image size per frame $64 \times 64 \times 1$) by using 4 SCLNs with 2, 2, 2, 2 strides on the spatial perceptual pathway and 4, 4, 4, 4 strides on the temporal perceptual pathway. The first SLCN uses two copies of X as the inputs into its spatial perceptual pathway and temporal perceptual pathway. Thus, beginning from the second SCLN, the generator takes the spatial and temporal perceptual pathway outputs of the previous SCLN as the input and encodes a $25 \times 4 \times 4 \times 128$ feature from the multi-attention weighing unit output of the fourth SCLN. Then, this encoded feature is further reduced to $1 \times 1 \times 4096$ by a fully connected layer and is then passed to another fully connected layer to reshape the encoded feature into a $4 \times 4 \times 256$ feature. Four upsampling blocks (Upsampling-Conv2D-LN) then use this reshaped feature to encode an image (i.e., the coarse tissue mask, the LGE-equivalent image or the diagnosis-related tissue segmentation image) $\in R^{64\times 64\times 1}$. Moreover, the generator also uses a dot layer to reduce the first dimension of the multi-attention weighing unit output from the first to the third SCLN and a skip connection that is the same as the U-Net to feed the corresponding upsampling block with the same feature map size.

The discriminator encodes the output of the generator of the corresponding phase and determines whether this output is consistent with the domain of its ground truth. All 5 convolutional layers have strides of 2. Note that the attention layer is added between the second convolutional layer and the third convolutional

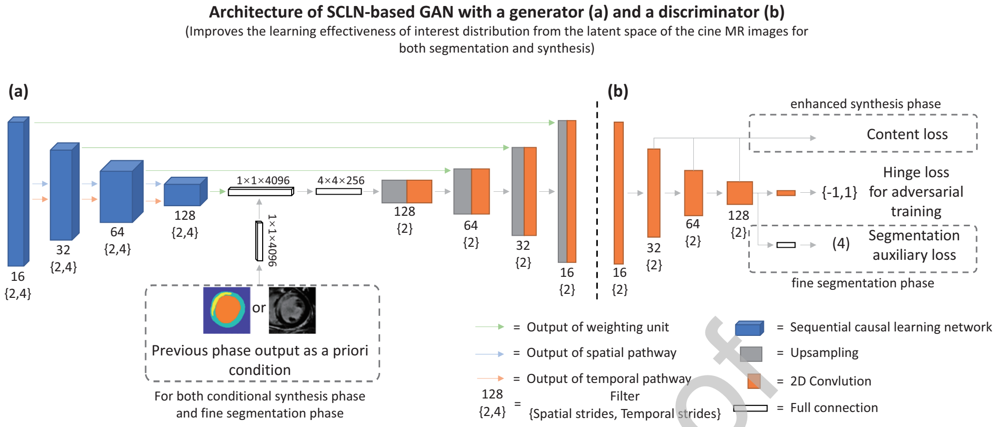
Architecture of SCLN-based GAN with a generator (a) and a discriminator (b)
(Improves the learning effectiveness of interest distribution from the latent space of the cine MR images for both segmentation and synthesis)
(a)
(b)
16 {2,4}
32 {2,4}
64 {2,4}
128 {2,4}
1×1×4096
4×4×256
1×1×4096
128 {2}
64 {2}
32 {2}
16 {2}
16 {2}
32 {2}
64 {2}
128 {2}
or
Previous phase output as a priori condition
For both conditional synthesis phase and fine segmentation phase
enhanced synthesis phase
Content loss
Hinge loss for adversarial training
{-1,1}
(4)
Segmentation auxiliary loss
fine segmentation phase
= Output of weighting unit
= Output of spatial pathway
= Output of temporal pathway
128 {2,4} = Filter {Spatial strides, Temporal strides}
= Sequential causal learning network
= Upsampling
= 2D Convlution
= Full connection

Figure 5: By integrating SCLN into the GAN architecture as the encoder of cine MR images in the generator, SCLN-based GAN improves the learning effectiveness of interest distribution from the latent space of cine MR images, thereby effectively improving the generating.

layer. These attention layers endow the discriminator with the ability to verify that highly detailed features in distant portions of the image are consistent with each other and to improve the discrimination performance.

In summary, the **advantage** of this SCLN-based GAN is an accurate encoding the interest dependencies from the latent space of cine MRI image.

5. Three progressive phases of PSCGAN

5.1. Phase I: priori generation GAN for coarse tissue mask generation

The priori generation GAN (Pri) is built with the same architecture as the SCLN-based GAN, as shown in Figure 6(a). It consists of a generator G_{Pri} and a discriminator D_{Pri}. This GAN generates a coarse tissue mask $\mathbf{M_{Pri}}$, which focuses on drawing the shape, contour and correct categories for the four classifications (scar, healthy myocardium, blood pool, and other pixels). This GAN does not seek a final result in one step but takes advantage of the shape, contour, and categories of this rough segmentation as a priori information to guide the next module to learn the attributes and distributions of the pixels.

Training of this generator uses multi-class cross-entropy loss. Although $\mathbf{M_{Pri}}$ contains four classes, the generator is treated as a single classification problem for the samples in one of these classes by encoding both the generator output and ground truth to one-hot vector classes. The generator can be formulated as follows:

$$\mathcal{L}_{G_{Pri}} = \sum_{n=1}^{N} \operatorname{mce}\left(G_{Pri}\left(X\right), \widetilde{I}_{Seg}\right) \tag{8}$$

$$\operatorname{mce} = -\frac{1}{N}\sum_{n=1}^{N}\left[\widetilde{I}_{Seg}\log M_{Pri} + \left(1-\widetilde{I}_{Seg}\right)\log\left(1-M_{Pri}\right)\right] \tag{9}$$

where $\widetilde{\mathbf{I}}_{\mathbf{Seg}}$ is the ground truth of $\mathbf{M}_{\mathbf{Pri}}$, and N = 4.

The discriminator training uses the adversarial loss $\mathcal{L}_{Adv}^{Pri}$, which adopts the recently developed hinge adversarial loss (Vaswani et al., 2017). This hinge adversarial loss maps the true sample to a range greater than 1 and maps the false sample to an interval less than -1. It better converges to the Nash equilibrium between the discriminator and generator, thus result in less mode collapsing and more stable training performance than other GAN losses Zhao et al. (2016). It can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{Adv}^{D_{Pri}} =& -\mathbb{E}_{(\widetilde{I}_{Seg})\sim p_{\text{data}}}[\min(0, -1 + D_{Pri}(\widetilde{I}_{Seg}))] \\ & - \mathbb{E}_{X\sim p_X}[\min(0, -1 - D_{Pri}(G_{Pri}(X)))] \\ L_{Adv}^{G_{Pri}} =& -\mathbb{E}_{X\sim p_X} D_{Pri}(G_{Pri}(X))\end{aligned} \tag{10}$$

5.2. *Phase II: conditional synthesis GAN for high-quality LGE-equivalent image synthesis*

The conditional synthesis GAN (Sys) consists of a generator G_{Sys} and a discriminator D_{Sys} to generate an LGE-equivalent image $\mathbf{I}_{\mathbf{Sys}}$. As shown in Figure 6(b), this GAN introduces the previously generated course tissue mask to guide the network training by modifying the SCLN-based GAN with a fully connected layer in the generator to concatenate the $1 \times 1 \times 4096$ feature and the mask, the output of which is then fed into the following fully connected layer and 4 upsampling blocks. Thus, this GAN builds a conditional joint mapping space between the segmentation and the synthesis to use the basic attributes and distributions (i.e., shape, contour, location, and categories) of the tissues to disentangle different tissue-feature learning in the cine MR images and allows the generator to perform accurate and detailed synthesis.

The generator uses the synthetic regularization loss $\mathcal{L}_{\mathbf{G}_{\mathbf{Sys}}}$ for the training. This loss incorporates an L2-regularization term and an overlapping group sparsity anisotropic operator (Peyré and Fadili, 2011) into the recently developed total

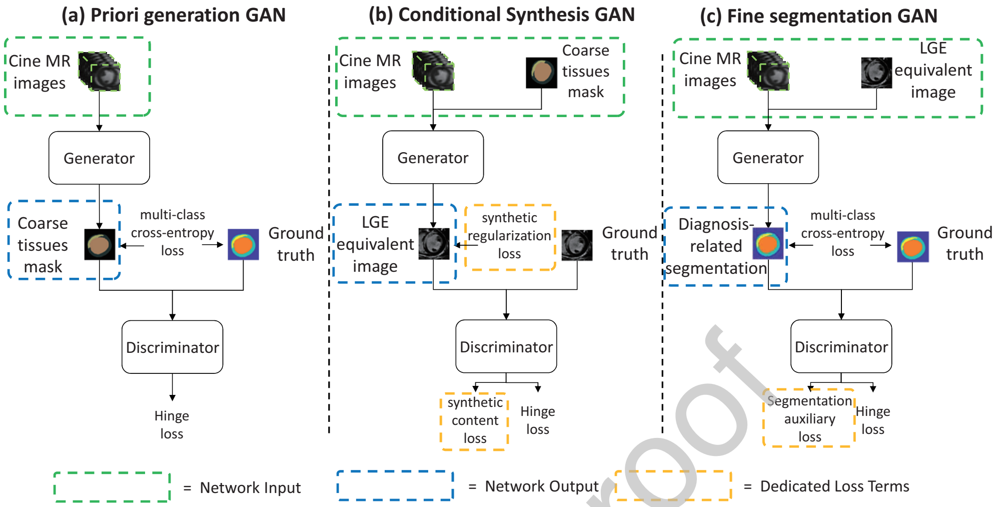

Figure 6: All GANs in the three phases leverage the adversarial training and dedicated loss terms to enhance the performance of synthesis and segmentation. Importantly, the conditional synthesis GAN and enhanced segmentation GAN leverage the output of the respective previous GANs to guide the training of the next GAN as part of its input.

variation loss to improve the quality of the synthesized image (Pumarola et al., 2018). The total variation loss has recently shown the ability to significantly reduce the noise in the synthesized image during image synthesis. L2-regularization is further incorporated into the total variation loss to measure the computation complexity and prevent overfitting by penalizing this complexity. The overlapping group sparsity anisotropic operator is further incorporated into the total variation loss. It takes into account group sparsity characteristics of image intensity derivatives, thereby avoiding staircase artifacts that erroneously consider smooth regions as piecewise regions (Peyré and Fadili, 2011). Concretely, this loss is formulated as follows:

$$\mathcal{L}_{\mathbf{G_{Sys}}} = \mathop{\mathbb{E}}_{\mathbf{I_{Sys}}\sim\mathbf{P_G}}[\frac{1}{2}\|\mathbf{I_{Sys}}\|_2^2+\nu(\phi(\mathbf{I_{Sys_{i+1,j}}}-\mathbf{I_{Sys_{i,j}}})+\phi(\mathbf{I_{Sys_{i,j+1}}}-\mathbf{I_{Sys_{i,j}}}))] \tag{11}$$

where i and j are the ith and jth pixel entry of $\mathbf{I_{Sys}}$, $\nu > 0$ is a regularization parameter, and $\phi(.)$ is overlapping group sparsity function. Overlapping group sparsity anisotropic operator is described as

$$\phi(u) = \sum_{i,j=1}^{n} \|u_{i,j,K}(:)\|_2 \tag{12}$$

$$\tilde{u}_{i,j,K} = \begin{bmatrix} u_{i-m_1,j-m_1} & u_{i-m_1,j-m_1+1} \\ u_{i-m_1+1,j-m_1} & u_{i-m_1+1,j-m_1+1} \end{bmatrix} \tag{13}$$

where K is the group size, $m_1 = \left\lfloor \frac{K-1}{2} \right\rfloor$ and $m_2 = \left\lfloor \frac{K}{2} \right\rfloor$.

The discriminator is trained using an adversarial loss term and a synthetic content loss term: 1) the synthesis adversarial loss $\mathcal{L}_{Adv}^{D_{Sys}}$ adopts the hinge adversarial loss and can be formulated as:

$$\begin{aligned} \mathcal{L}_{Adv}^{D_{Sys}} =& -\mathbb{E}_{(\widetilde{I}_{Sys})\sim p_{\text{data}}}[\min(0, -1 + D_{Seg}(\widetilde{I}_{Sys}))] \\ & - \mathbb{E}_{X\sim p_X}[\min(0, -1 - D_{Sys}(G_{Sys}(X|M_{Pri})))] \\ L_{Adv}^{G_{Sys}} =& -\mathbb{E}_{X\sim p_X} D_{Sys}(G_{Sys}(X|M_{Pri})) \end{aligned} \tag{14}$$

where $\widetilde{\mathbf{I}}_{\mathbf{Sys}}$ is the ground truth (i.e, LGE image).

2) the synthetic content loss $\mathcal{L}_{Cont}^{Sys}$ is specially designed to use feature maps of the 2nd, 3rd and 4th convolution layers outputted from discriminator to evaluate $\mathbf{I}_{\mathbf{Sys}}$ by comparing it to its ground truth $\widetilde{\mathbf{I}}_{\mathbf{Sys}}$. This multiple feature map evaluation allows the discriminator to discriminate the image in terms of both the general detail content and higher detail abstraction during the activation of the deeper layers, thereby improving the discriminator performance (Johnson et al., 2016). It is defined as follows:

$$\mathop{\mathbb{E}}_{I_{Sys}\sim\mathbb{P}_{data}}[\frac{1}{W_i H_i}\sum_{x=1}^{W_i}\sum_{y=1}^{H_i}(D_{Sys}^{Conv_i}(\widetilde{I}_{Sys})_{x,y} - D_{Sys}^{Conv_i}(G_{Sys}(X|M_{Pri})_{x,y}))^2] \tag{15}$$

where $D_{Sys}^{Conv_i}$ is the feature map and W_i and H_i obtained by the ith convolution layer (after activation).

In summary, the **advantages** of the conditional synthesis GAN are as follows: 1) the coarse tissue mask is used as an a priori condition to guide the accurate synthesis of the tissues, 2) the synthetic regularization loss is used to reduce the image noise during synthesis, and 3) the synthetic content loss is used to improve the detail restoration in the image synthesis.

5.3. Phase III: enhanced segmentation GAN for accurate diagnosis-related tissues segmentation

The enhanced segmentation GAN (Seg) consists of a generator G_{Seg} and a discriminator D_{Seg} to generate an accurate diagnosis-related tissue segmentation image $\mathbf{I_{Seg}}$, as shown in Figure 6(c). Compared to the basic SCLN-based GAN, this GAN has following two differences: 1) it adds a fully connected layer into the generator at the same position as that of the conditional synthesis GAN to introduce the synthesized image output from phase II as a condition to guide the segmentation. The synthesized image already includes all detailed textures of the tissues, which effectively aids the fine classification of the tissue boundary pixels, and 2) it adds a linear layer at the end of the discriminator to regress the size (number of pixels) of the 4 different segmentation categories at the end of the discriminator to perform a self-supervised segmentation auxiliary loss. This self-supervised loss prevents the discriminator from only judging the segmented image based on the segmentation shape, causing the discriminator to extract a compensate feature from the input image to improve its discrimination performance. Concretely, the generator with multi-class cross-entropy loss and the discriminator with segmentation adversarial loss are formulations as follows:

$$\begin{aligned}
\mathcal{L}_{G_{Seg}} &= \sum_{n=1}^{N} \mathrm{mce}\left(G_{Seg}\left(X|I_{Sys}\right), \widetilde{I}_{Seg}\right) \\
\mathcal{L}_{Adv}^{D_{Seg}} &= -\mathbb{E}_{(\widetilde{I}_{Seg})\sim p_{\text{data}}}[\min(0, -1 + D_{Seg}(\widetilde{I}_{Seg}))] \\
&\quad - \mathbb{E}_{X\sim p_X}[\min(0, -1 - D_{Seg}(G_{Seg}(X|I_{Sys})))] \\
L_{Adv}^{G_{Seg}} &= -\mathbb{E}_{X\sim p_X} D_{Seg}(G_{Seg}(X|I_{Sys}))
\end{aligned} \tag{16}$$

The discriminator with self-supervised segmentation auxiliary loss is formulation as follows:

$$\mathcal{L}_{Seg}^{Aux} = \mathbb{E}_{\widetilde{I}_{Seg}\sim P_{data}} ||D_{Seg}^{Aux}(Si|\widetilde{I}_{Seg}) - D_{Seg}^{Aux}(Si|G_{Seg}(X|I_{Sys})))||_1 \tag{17}$$

where $Si = \sum_{n=1}^{4}(Si_1, Si_2, Si_3, Si_4)$ is the size of the 4 segmentation categories of pixels in the image outputted from the linear layer of the discriminator D_{Seg}^{Aux}.

In summary, the **advantages** of the enhanced segmentation GAN are as follows: 1) the boundaries of tissues within synthesized images are used to guide the tissues boundary segmentation and 2) the self-supervised segmentation auxiliary loss is used to improve the segmentation adversarial.

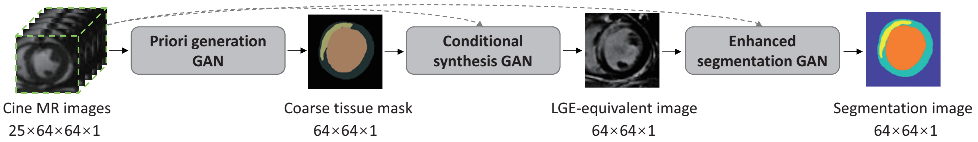

Figure 7: PSCGAN cascade three GANs and connects them by taking the output of the previous GAN as an input of the next GAN.

6. Materials and Implementation

6.1. Materials

A total of 280 (230 IHD and 50 normal control) patients with short-axis cine MR images were selected. Cardiac cine MR images were obtained using a 3-T MRI system (Verio, Siemens, Erlangen, Germany). Retrospectively gated balanced steady-state free-precession nonenhanced cardiac cine images with 25 reconstructed phases were acquired (repetition time/echo time, 3.36 msec/1.47 msec; field of view, 286×340 mm^2; matrix, 216×256; average temporal resolution, 40 msec). LGE MRI was performed in the same orientations and with the same section thickness using a two-dimensional segmented, fast low-angle shot, phase-sensitive inversion recovery sequence 10 minutes after intravenous injection of a gadolinium-based contrast agent (Magnevist, 0.2 mmol/kg; Bayer Healthcare, Berlin, Germany). Moreover, a network with heart localization layers, as described in (Xu et al., 2017), was used to automatically crop both cine MR images and LGE images to 64×64 region-of-interest sequences, including the left ventricle. Furthermore, the cropped cine and LGE images were registered at the end-diastole phase.

6.2. Ground truth

The ground truth of the LGE-equivalent image is the real LGE images . The ground truth of the diagnosis-related tissue segmentation image is an LGE segmented image that includes the contours of the healthy myocardium, scar, and blood pool. These contours were manually delineated on the LGE MRI by a radiologist (N.Z., with 7 years of experience in cardiovascular MRI) from the LGE image. All manual segmentations were reviewed by another expert (L.X., with 10 years of experience in cardiovascular MRI), and in cases of disagreement, a consensus was reached.

6.3. Implementation detail

The PSCGAN randomly selected 3/4 of the patients for training and the remaining 1/4 (70) patients were used for independent testing. All three GANs were trained using an ADAM solver (Kingma and Ba, 2014) with a batch size of 1 and an initial learning rate of 0.001. For every 2 optimization steps of the discriminator, we performed a single optimization step for the generator. Layer normalization (Ba et al., 2016) and LeakyReLU activation (Goodfellow et al., 2016) were used both in the generators and the discriminators. The pixel values were normalized to [-1, 1].

6.4. Algorithm summary

Figure 7 indicates that PSCGAN connect three GANs by taking the output of the previous GAN as an input of the next GAN. Each GAN includes a generator and a discriminator. All discriminators are used only during adversarial training.

- Priori generation GAN inputs the 2D+T cine MR images $\mathbf{X} \in \mathbb{R}^{H \times W \times T \times C}$, where $H = W = 64$ are the height and width of each temporal frame, $T = 25$ is a temporal step, $C = 1$ is the number of channels. This GAN outputs coarse tissue masks of $64 \times 64 \times 1$. When adversarial training, the generator of this GAN inputs 2D+T cine MR images and outputs coarse tissue masks. The discriminator of this GAN inputs coarse tissue masks and the corresponding ground truth is $64 \times 64 \times 1$. This discriminator outputs 1×4 probability values.

- Conditional synthesis GAN inputs a combination of coarse tissue masks of $64 \times 64 \times 1$ and cine MR images of $25 \times 64 \times 64 \times 1$. This GAN outputs outputs LGE-equivalent images of $64 \times 64 \times 1$. During the adversarial training, the generator of this GAN inputs a combination of coarse tissue masks, and cine MR images, and ouputs LGE-equivalent images. The discriminator of this GAN inputs LGE-equivalent images and the corresponding ground truth of $64 \times 64 \times 1$. This discriminator outputs 1×1 probability values.

- Enhanced segmentation GAN inputs the combination of LGE-equivalent images of $64 \times 64 \times 1$ and cine MR images of $25 \times 64 \times 64 \times 1$. This GAN outputs diagnosis-related tissue segmentation images of $64 \times 64 \times 1$. During the adversarial training, the generator of this GAN inputs a combination of LGE-equivalent images and cine MR images, and outputs diagnosis-related

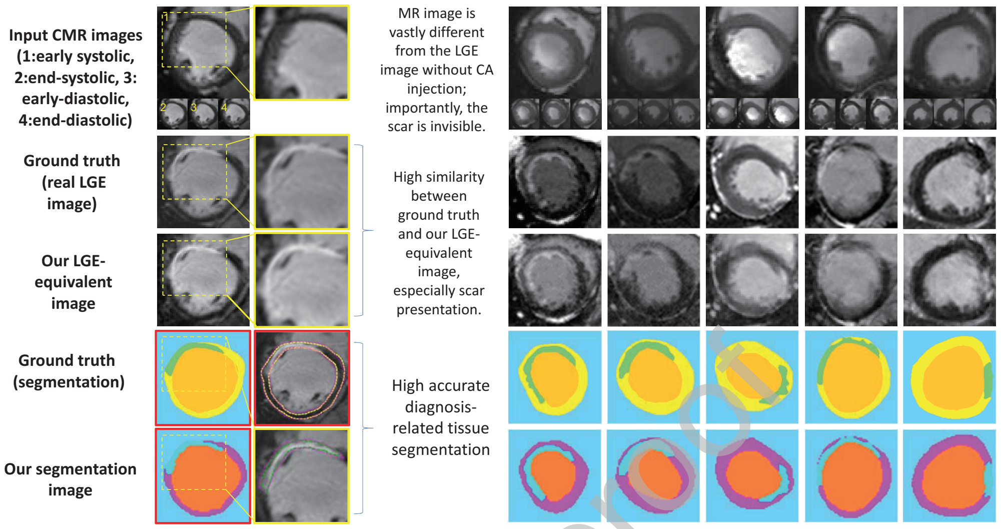
Input CMR images (1:early systolic, 2:end-systolic, 3: early-diastolic, 4:end-diastolic)
1
2
3
4
MR image is vastly different from the LGE image without CA injection; importantly, the scar is invisible.
Ground truth (real LGE image)
Our LGE-equivalent image
High similarity between ground truth and our LGE-equivalent image, especially scar presentation.
Ground truth (segmentation)
Our segmentation image
High accurate diagnosis-related tissue segmentation

Figure 8: PSCGAN synthesize high-quality LGE-equivalent images and produces accurate diagnosis-related tissue segmentation images. In LGE-equivalent images, the scar (dashed box, the high contrast area in LV wall) has a clear and accurate presentation when compared to the real LGE image. Note that this high contrast area is invisible in cine MR images without CA injection. In diagnosis-related tissue segmentation images, the segmented scar (cyan 🟦), health myocardium (purple 🟪), and blood pool (orange 🟧) from our method are highly consistent with the ground truth in terms of shape, location, and size.

tissue segmentation images. The discriminator of this GAN inputs LGE-equivalent images and the corresponding ground truth of $64 \times 64 \times 1$. This discriminator outputs 1×4 probability values, and 1×4 vectors.

Note that the $64 \times 64 \times 1$ coarse tissue masks and segmented images are categorical data, which are quickly converted to and from $64 \times 64 \times 4$ one-hot data during adversarial training.

6.5. *Metrics*

Our network evaluates its performance in two aspects: 1) clinical metrics and 2) imageology metrics. In clinical metrics, our network evaluates the scar size, the segment-level scar localization (16-segment model), the MI ratio (scar pixels/ healthy myocardium pixels), and the transmurality. All these metrics compare the results of our diagnosis-related tissue segmentation image with the results of the ground truth by using the correlation coefficient, Bland-Altman analysis (Altman and Bland, 1983), sensitivity, specificity and positive and negative predictive values (PPV and NPV). In imageology metrics, our network compares our

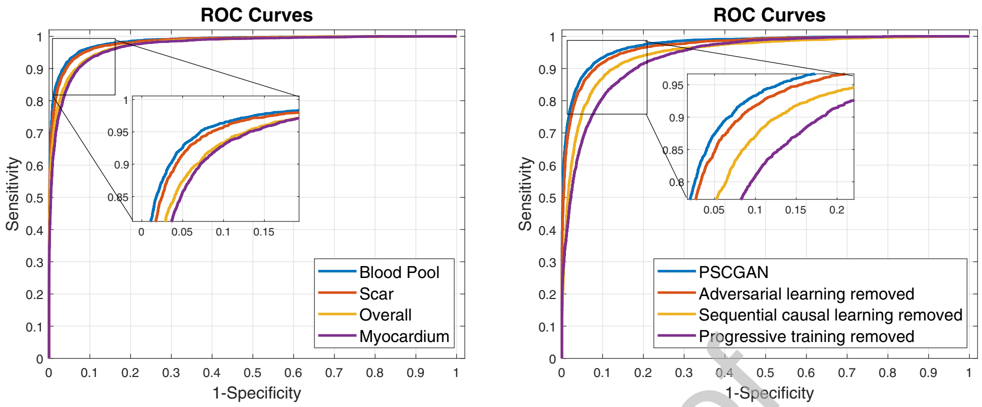

Figure 9: PSCGAN generated an accurate diagnosis-related tissue segmentation image. Furthermore, each technologically innovative component in the PSCGAN effectively improve the segmentation accuracy.

segmented image with the ground truth by calculating the accuracy, sensitivity, specificity, and Dice coefficient. The network also compares the LGE-equivalent image with the LGE image (ground truth) by calculating the structural similarity index (SSIM) (Wang et al., 2004), peak signal-to-noise ratio (PSNR) (Welstead, 1999), and normalized root-mean-squared error (NRMSE).

7. Experiments and Results

Comprehensive experiments indicated that the PSCGAN synthesize high-quality LGE equivalent image and accurately segments all diagnosis-related tissues. PSCGAN achieved an NRMSE of 0.14 when comparing the LGE equivalent image to ground truth and achieved 97%, 96%, and 97% segmentation accuracy when comparing the clinicians manual segmentation of the scar, healthy myocardial tissues, and blood pools, respectively. The correlation coefficient between the scar ratio obtained from PSCGAN and that from the current clinical workflow was 0.96. These results demonstrated that PSCGAN could perform full diagnosis-related tissue observation and segmentation, thereby obtaining highly accurate diagnosis metrics in a real clinic setting.

7.1. High-quality LGE-equivalent image synthesis and accurate diagnosis-related tissues segmentation

7.1.1. Imageology metrics

Table 1 and Figure 8 indicate that PSCGAN were able to synthetize high-quality LGE-equivalent images, which were almost identical to the LGE image based on CA injection, in terms of the imageology metrics . It achieved an SSIM of 0.78±0.10, a PSNR of 23.03±1.42, and an NRMSE of 0.11±0.05. Moreover, PSCGAN achieved an average SSIM of 0.76±0.18, a PSNR of 23.17±1.60, and an NRMSE of 0.10±0.09 when using the 10-fold random cross-validation. Note that higher values for SSIM and PSNR and lower values for NRMSE indicated better performance.

Table 1, Figure 8, and Figure 9 shows that PSCGAN accurately segmented IHD scars, healthy myocardium and blood pools in terms of the imageological metrics. Our method achieved an overall pixel segmentation accuracy of 97.17% with a sensitivity of 91.68% and a specificity of 98.53%. In particular, the accuracy of the scar segmentation is 97.13%, that of the healthy myocardium segmentation is 96.34% and that of the blood pool segmentation is 97.97%. PSCGAN obtained Dice coefficients of 0.93 for the scar tissue, 0.90 forthe healthy myocardial tissue, and 0.93 for the blood pools. Moreover, when using the 10-fold random cross-validation, our method achieved an overall pixel segmentation accuracy of 97.11% with a sensitivity of 91.24% and a specificity of 98.67%. In particular, the accuracy of the scar segmentation is 96.94%, that of the healthy myocardium segmentation is 96.37% and that of the blood pool segmentation is 98.01%. PSCGAN obtained Dice coefficients of 0.90 for the scar tissue, 0.91 for the healthy

Table 1: The PSCGAN achieved accurate diagnosis-related tissues segmentation image and high-quality LGE-equivalent image synthesis in terms of imageology metrics

Accurate diagnosis-related tissues segmentation image

	Accuracy	Sensitivity	Specificity	Dice coefficient
Overall	97.17(0.48)%	91.68%	98.53%	0.918(0.17)
Scar	97.13(0.23)%	90.84%	98.48%	0.932(0.11)
Healthy myocardium	96.34(0.51)%	91.07%	99.11%	0.908(0.19)
Blood pool	97.97(0.44)%	91.84%	98.36%	0.936(0.15)

High-quality LGE-equivalent image synthesis

SSIM	NRMSE	PSNR
0.78(0.10)	0.11(0.05)	23.03(1.42)

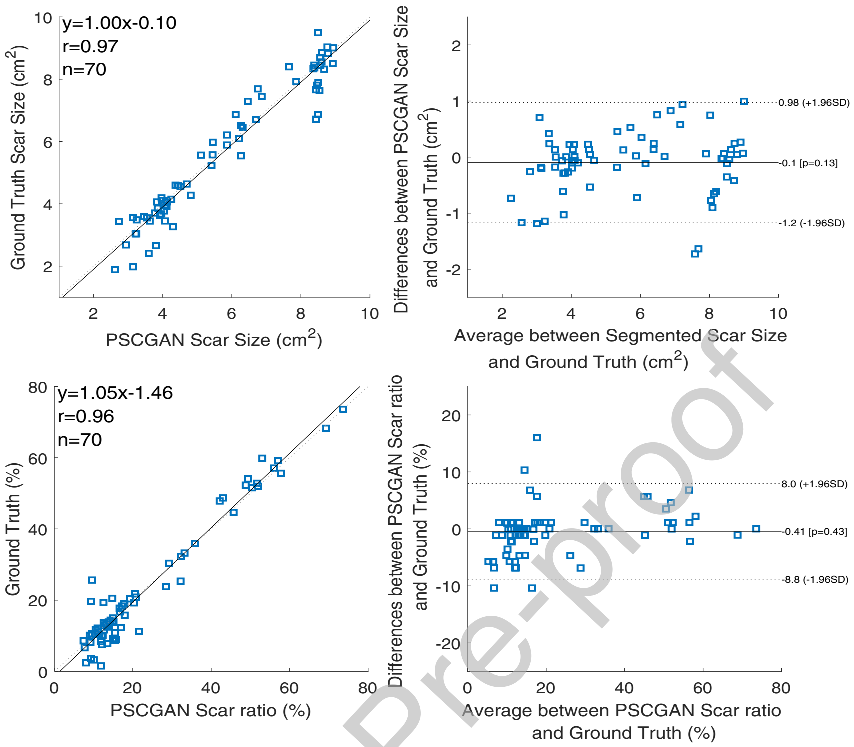

Figure 10: PSCGAN calculated scar sizes and scar ratios highly consistent with those from the current clinical workflow as shown by comparisons with Bland-Altman analysis.

myocardial tissue, and 0.93 for the blood pools.

7.1.2. Clinical metrics

The experimental results also show that PSCGAN can provide radiologists with the same clinical metrics for diagnosis as current clinical workflows, as shown in Figure 10 and Table 3. When compared to the ground truth, the PSCGAN achieved a correlation coefficient of 0.97 and -0.1 (0.98,-1.2) cm^2 for the corresponding biases (limits of agreement) in scar size, a sensitivity of 85.27% and a specificity of 97.47% in the segment-level scar localization, a correlation coefficient of 0.96 and 0.41 (8.0, -8.8)% for the corresponding biases (limits of agreement) in scar ratio, and a sensitivity of 86.95% and a specificity of 97.87% in scar transmurality. Moreover, when using the 10-fold random cross-validation, the PSCGAN achieved a correlation coefficient of 0.95 in scar size, a sensitivity of 84.80% and a specificity of 97.67% in the segment-level scar localization,

Table 2: Clinical metrics obtained by PSCGAN are highly consistent with those obtained from the current clinical workflow.

	Sensitivity	Specificity	AUC
Scar segment-level localization	85.27%	97.47%	0.90
Scar transmurality	86.95%	97.87%	0.91
	PSCGAN	Ground truth	Pearson's r (P-value)
Scar size (cm^2)	7.37 ± 2.17	5.64 ± 1.93	0.97 (0.24)
Scar ratio(%)	29.10±19.73	25.31±17.62	0.96 (0.11)

a correlation coefficient of 0.94, in scar ratio, and a sensitivity of 82.61% and a specificity of 98.17% in scar transmurality.

7.2. Advantage of the generative adversarial learning

Figure 6 and Table 3 indicates that generative adversarial learning improves the performance of both the segmentation and the synthesis. Among them, the improvement of synthesis is particularly obvious. The generative adversarial learning of PSCGAN improved overall segmentation accuracy by 1.2%, the SSIM by 0.23, and the pearsons r of scar size by 0.02 compared to a network with adversarial learning removed, which only uses an SCLN-based generator with parallel output for segmentation and synthesis. Moreover, PSCGAN improved overall segmentation accuracy by 0.94%, the SSIM by 0.21, and the pearsons r of scar size by 0.02 when using the 10-fold random cross-validation. This improved performance fully proves that generative adversarial learning using game theory enables the

Table 3: Each technological innovation component in PSCGAN has effectively improved the its performance.

	Accuracy of overall segmentation image	SSIM of CA-free enhancement image	Pearsons r of scar size
PSGAN	**97.17(0.48)**%	**0.78(0.10)**	**0.97**
Adversarial learning removed	95.92(0.57)%	0.55(0.21)	0.95
Progressive training removed	94.91(0.59)%	0.61(0.19)	0.93
Sequential causal learning removerd (3DConv)	95.13(0.50)%	0.64(0.17)	0.96

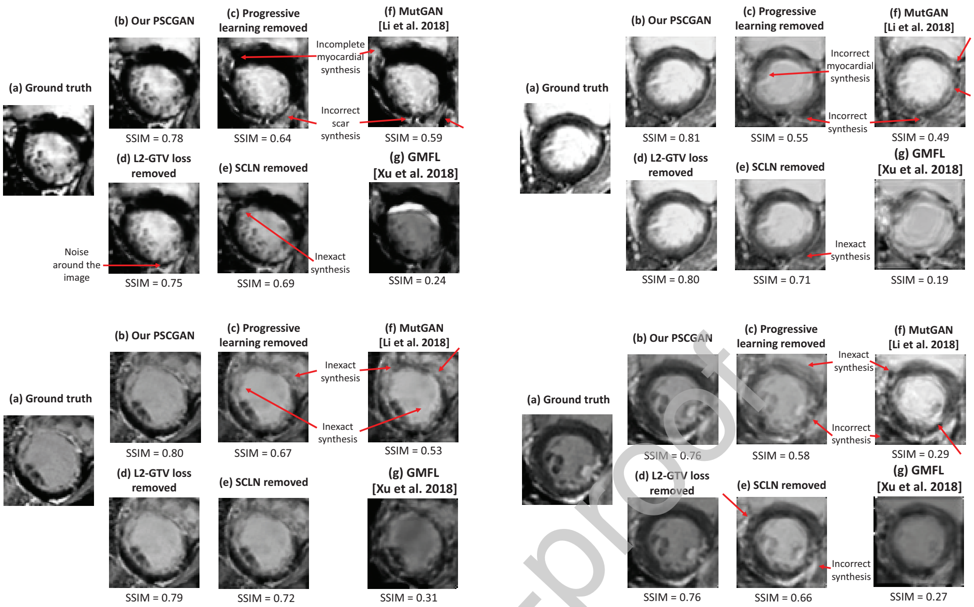

Figure 11: Each technologically innovative component in the PSCGAN effectively improves LGE-equivalent images quality.

learning of better representations from a latent space of data distribution, thereby optimizing the segmentation contours and enhancing the fine synthesis details.

7.3. Advantage of the progressive training framework

Figures 9 and 11 and Table 3 indicate that the progressive framework of the PSCGAN significantly improves the training stability while improving the learning efficiency and accuracy in both the segmentation and the synthesis. The PSCGAN improved the overall segmentation accuracy by 2.2%, the SSIM by 0.17, and the pearsons r of scar size by 0.04 compared with a network with the progressive framework removed that produced a parallel output of segmentation and synthesis using one generator (G_{pri}) and one discriminator (D_{pri}). The PSCGAN improved the overall segmentation accuracy by 1.92%, the SSIM by 0.11, and the pearsons r of scar size by 0.03 when using the 10-fold random cross-validation and progressive framework removed network. The standard deviation of the segmentation accuracy of the full PSCGAN was also reduced by 0.11% compared to the network with the framework removed, while the standard deviation of the SSIM was reduced by 0.09. Furthermore, the progressive framework also reduced the difference between the segmentation results from the ground truth and those

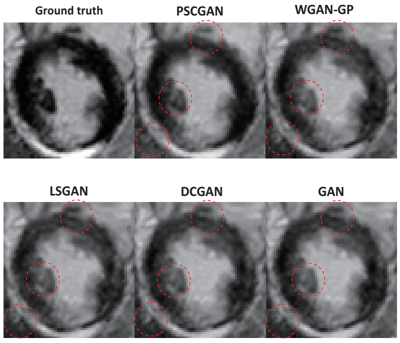

Figure 12: The hinge adversarial loss term in the PSCGAN achieved the best performance in LGE-equivalent image synthesis.

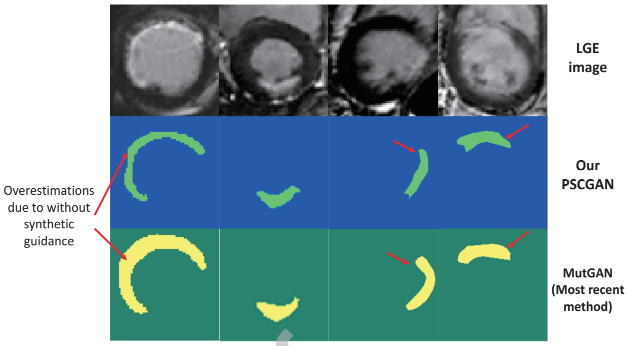
LGE image
Our PSCGAN
MutGAN (Most recent method)
Overestimations due to without synthetic guidance

Figure 13: PSCGAN obviously correct the overestimation and boundary error issues in existing state-of-the-art scar segmentation methods.

from the LGE-equivalent images (0.09% in the PSCGAN and 1.20% in the progressive framework-removed version). All these improvements proved that our progressive framework created joint mappings that successively augmented the tissue mask and LGE-equivalent images in the synthesis and segmentation training. These joint mappings successfully exploited the commonalities between the LGE-equivalent images and the diagnosis-related segmentation images, thereby avoiding interference between the conditional probability distribution of the generative model-based synthetic task and the decision function of the discriminative model-based segmentation task.

Table 4: SCLN outperforms recent time-series image learning methods, and each component in the SCLN effectively improves performance.

	Full SCLN	Spatial perceptual pathway only	Temporal perceptual pathway only	multi-attention weighing removed	ConvLSTM	3DConv +LSTM
Accuracy (Overall)	**97.17%**	73.61%	89.42%	96.72%	95.97%	96.47%
SSIM	**0.78**	0.48	0.57	0.74	0.71	0.70
Pearsons r (scar size)	**0.97**	0.71	0.83	0.94	0.91	0.93

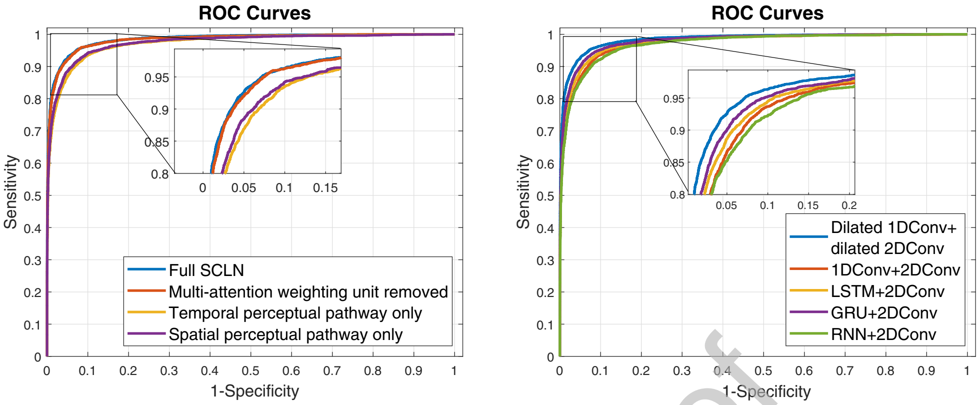

Figure 14: The two-stream pathways and the weighing unit in the SCLN effectively improve segmentation accuracy, as does multi-scale, causal dilated convolution.

7.4. *Advantage of the sequential causal learning network*

Figures 9, 11, and 14 and Tables 3 and 4 indicate that the SCLN effectively improved both segmentation accuracy and synthesis quality. Compared with the current 2D+T time-series learning methods, SCLN improved the segmentation accuracy, the SSIM and the pearsons r of scar size by 2.14%, 0.14 and 0.01, respectively, compared to Conv3D, by 1.13%, 0.07 and 0.06, respectively, compared to ConvLSTM, and by 0.73%, 0.08 and 0.04, respectively, compared to 3DConv+LSTM. This is because SCLN creates a multi-scale, two-stream extractor to match spatial and temporal dependencies in time-series image learning, thereby avoiding the interference between these two dependencies during

Table 5: Synthetic regularization loss effectively improved the quality of the LGE-equivalent images. Segmentation auxiliary loss also effectively improved the accuracy of the diagnosis-related tissue segmentation images.

PSCGAN		Synthetic regularization loss removed	
SSIM	PSNR	SSIM	PSNR
0.78(0.10)	**23.03(1.42)**	0.77(0.12)	21.50(2.07)
PSCGAN		Segmentation auxiliary loss removed	
Accuracy(overall)		Accuracy(overall)	
97.17(0.48)%		97.04(0.58)%	

learning, and a multi-attention weighing unit is used to select the task-specific dependencies between and within the spatial and temporal dependencies. Particularly, the experimental results also indicate that each component of the SLCN effectively improved performance, especially that of synthesis, as shown in Figure 14 and Table 4. Compared with the spatial perceptual pathway-alone version, the temporal perceptual pathway-alone version, and the multi-attention weighing unit-removed version, the SCLN shows improvements of 23.56%, 7.75%, and 0.45%, respectively, in segmentation accuracy, improvements of 0.30, 0.21, and 0.04, respectively, in SSIM, and improvements of 0.26, 0.14, and 0.03, respectively, in the Pearsons r of the scar size. Furthermore, Figure 14 and Table 4 also indicate that, within the SCLN, multi-scale 2D causal dilated convolution + 1D causal dilated convolution drive both the spatial perceptual pathway and the temporal perceptual pathway to achieve better performance. Compared with the other temporal information and spatial information separating learning methods, SCLN improved the segmentation accuracy and the SSIM by 2.65% and 0.08, respectively, compared to 2DConv+1DConv, by 1.95% and 0.05, respectively, compared to LSTM+2DConv, by 1.87% and 0.03, respectively, compared to GRU+2DConv, and by 4.06% and 0.17, respectively, compared to RNN+2DConv. This is because multi-scale, causal dilated convolution successfully handles the high local variations of pixels in the cine MR images by changing the dilation ratio to extract both long-range and short-range spatial and temporal dependencies.

The cases where our method fails are illustrated in Figure. 15, and mainly focus on the inaccurate synthesis and segmentation of scars. The main reason for these failures may be because our method only relies on the cine MR images for the spatiotemporal representation learning of the heart. The spatiotemporal representation of the heart is a very complex 3D change in both kinematics and morphology. Although cardiac cine MR images are the most effective and widely protocol for imaging the beating heart, they are single short-axis images and are insufficient for presenting a complete spatiotemporal representation of the 3D swirl and spiral of the muscle cells in the heart. Nevertheless, this problem can be improved by introducing extra modality images (such as T2WI images) and extra view images (such as long-axis images) in the further work.

7.5. Advantage of synthetic regularization loss and segmentation auxiliary loss

Figure 11 and Table 5 indicate that synthetic regularization loss improved the quality of the synthesized image, especially in terms of PSNR. Synthetic regularization loss improved the PSNR by 1.8 compared to the network with the loss

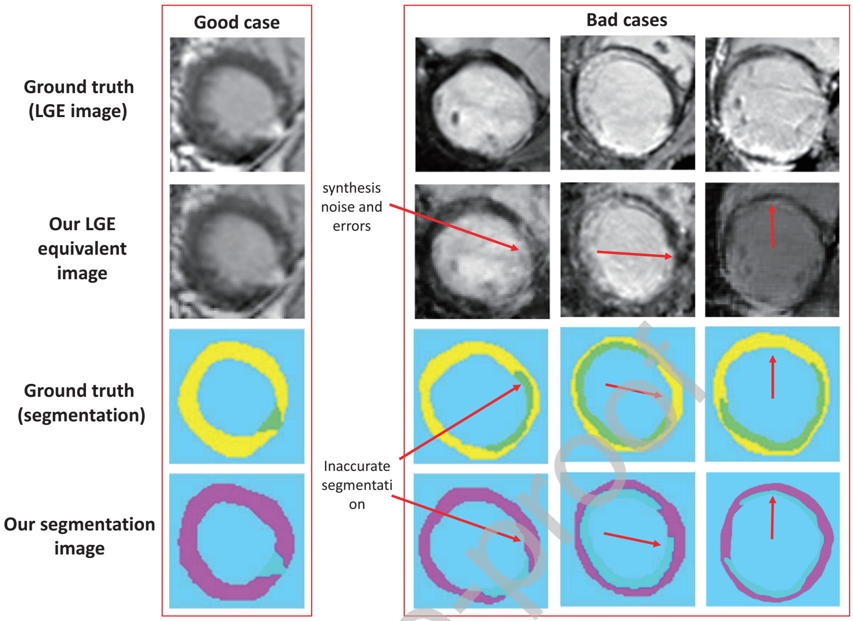

Figure 15: Visual examples of the synthesis and segmentation, including both good case and bad cases (red arrows). Note that segmented scars appear as green and cyan areas in our method and the ground truth, respectively. The segmented myocardium appear as yellow and purple areas in our method and the ground truth, respectively.

term removed. This is because synthetic regularization loss builds a group sparsity structure that has a natural grouping of its components and the components within a group. Thus, this loss reduces the degrees of freedom in the total variation during noise optimization, thereby leading to better synthesis performance. Moreover, Table 5 indicates that segmentation auxiliary loss improved the overall segmentation accuracy by 0.13% compared with the version with this loss term removed. This is because the segmentation auxiliary loss adds additional tissue size information to the discriminator, thereby motivating the network to learn more aspects of the distribution of the segmented images to improve the performance of the network. In addition, the experimental results indicated that hinge adversarial loss has the overall best performance when compared with other, recently developed adversarial losses. In terms of segmentation, hinge adversarial loss term achieved the highest accuracy (97.17%), which was the same as that of WGAN-GP loss (Gulrajani et al., 2017) and LSGAN loss terms (Mao et al., 2017). In terms of synthesis, the hinge adversarial loss term achieved the highest SSIM

(0.78), and the WGAN-GP loss term achieved the second highest SSIM (0.76).

7.6. Comparison with other state-of-the-art methods

The PSCGAN represent the first networks to combine CA-free IHD-diagnosing image synthesis and segmentation technologies, produced a greater number of diagnosis metrics and yielded higher IHD segmentation and diagnosis accuracies than existing state-of-the-art methods (Zhang et al., 2019; Bleton et al., 2015; Xu et al., 2017; Popescu et al., 2016; Xu et al., 2018a), as shown in Table 6. Concretely, PSCGAN improved scar segmentation accuracy 0.36%-12.74% compared to the other methods. PSCGAN obviously correct the overestimation and boundary error issues in existing state-of-the-art scar segmentation methods, as shown in Figure 13, by leveraging the textures and edges in LGE-equivalent images as priori conditions and by also leveraging the novel segmentation auxiliary loss terms. Moreover, PSCGAN successfully synthesized LGE-equivalent images. Note that some existing segmentation methods can be used mechanically for the synthesis

Table 6: PSCGAN achieved more diagnosis metrics and higher segmentation and diagnosis accuracy than existing state-of-the-art methods in IHD diagnosis and segmentation.

	Seg/Sys	Accuracy (Scar)	Accuracy (Overall)	SSIM	Pearson's r for scar ratio
PSGAN	**Sys /Multi-Seg**	**97.13%**	**97.17%**	**0.78**	**0.96**
(Xu et al., 2018a)	only scar Seg	96.77%	NaN (94.60%)	0.59*	NaN (0.93)
(Zhang et al., 2019)	only scar Seg	95.03%	NaN (92.37%)	0.31*	NaN (0.84)
(Xu et al., 2017)	only scar Seg	94.93%	NaN (92.51%)	0.31*	NaN (0.83)
(Popescu et al., 2016)	only scar Seg	86.47%	-	-	-
(Bleton et al., 2015)	only scar Seg	84.39%	-	-	-

NaN(.) means that this method can only estimate this index after the radiologist manually segments the endocardium and epicardium.

* means that the framework of this method can be used to synthesize LGE-equivalent image.

- means that this method is completely incapable of estimating this index.

task due to having the same input and output formats. PSCGAN achieved the highest SSIM values and improved the image quality in terms of scar presentation, boundary clarity, texture accuracy, and noise control, as shown in Figure 11. This is because the specially designed progressive framework, the SCLN, and the synthetic regulation loss terms built accurate spatiotemporal representations of the cine MR images for each pixel of the LGE image. Importantly, the PSCGAN produced credible diagnosis metrics that cannot be produced by all existing IHD diagnosis and segmentation methods, such as scar ratio. This is because PSCGAN enable the segmentation of all diagnosis-related tissues used for credible diagnosis metrics, rather than only scar-based metrics produced by existing binary segmentation methods.

8. Conclusion

For the first time, a progressive sequential causal GAN was used as a successful one-stop IHD-diagnosing CA-free technology to simultaneously synthesize an LGE-equivalent image and segment all diagnosis-related tissues from cine MR images. The PSCGAN were run using data from 180 subjects and yielded an SSIM for the synthesized image of 0.78, a scar pixel classification accuracy of 97.13%, and an overall, diagnosis-related tissue segmentation accuracy of 97.17%. These results demonstrate that the PSCGAN can be an efficient and accurate clinical tool for the substantial standardization of IHD diagnosis and can avoid all of the emerging toxicity concerns associated with CA.

References

Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. Journal of the Royal Statistical Society: Series D (The Statistician) 32, 307–317.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Beckett, K.R., Moriarity, A.K., Langer, J.M., 2015. Safe use of contrast media: what the radiologist needs to know. Radiographics 35, 1738–1750.

Bijnens, B., Claus, P., Weidemann, F., Strotmann, J., Sutherland, G.R., 2007. Investigating cardiac function using motion and deformation analysis in the setting of coronary artery disease. Circulation 116, 2453–2464.

Bleton, H., Margeta, J., Lombaert, H., Delingette, H., Ayache, N., 2015. Myocardial infarct localization using neighbourhood approximation forests, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 108–116.

Duchateau, N., De Craene, M., Allain, P., Saloux, E., Sermesant, M., 2016. Infarct localization from myocardial deformation: prediction and uncertainty quantification by regression from a low-dimensional space. IEEE transactions on medical imaging 35, 2340–2352.

Fox, C.S., Muntner, P., Chen, A.Y., Alexander, K.P., Roe, M.T., Cannon, C.P., Saucedo, J.F., Kontos, M.C., Wiviott, S.D., 2010. Use of evidence-based therapies in short-term outcomes of st-segment elevation myocardial infarction and non–st-segment elevation myocardial infarction in patients with chronic kidney disease. Circulation 121, 357–365.

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. volume 1. MIT press Cambridge.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, pp. 5767–5777.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., Belongie, S., 2017. Stacked generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5077–5086.

Ingkanisorn, W.P., Rhoads, K.L., Aletras, A.H., Kellman, P., Arai, A.E., 2004. Gadolinium delayed enhancement cardiovascular magnetic resonance correlates with clinical measures of myocardial infarction. Journal of the American College of Cardiology 43, 2253–2259.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. arXiv preprint .

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer. pp. 694–711.

Kali, A., Cokic, I., Tang, R.L., Yang, H.J., Sharif, B., Marbán, E., Li, D., Berman, D., Dharmakumar, R., 2014. Determination of location, size and transmurality of chronic myocardial infarction without exogenous contrast media using cardiac magnetic resonance imaging at 3t. Circulation: Cardiovascular Imaging , CIRCIMAGING–113.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 .

Kingma, D.P., Ba, J.L., 2014. Adam: Amethod for stochastic optimization, in: Proc. 3rd Int. Conf. Learn. Representations.

Ledesma-Carbayo, M.J., Kybic, J., Desco, M., Santos, A., Suhling, M., Hunziker, P., Unser, M., 2005. Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. IEEE transactions on medical imaging 24, 1113–1126.

Leiner, T., 2019. Deep learning for detection of myocardial scar tissue: Goodbye to gadolinium. Radiology 291.

Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 .

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 2813–2821.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .