Segmentation and Quantification of Infarction without Contrast Agents via Spatiotemporal Generative Adversarial Learning

Chenhu Xu\textsuperscript{a}, Joanne Howey\textsuperscript{a}, Pavlo Ohorodnyk\textsuperscript{a}, Mike Roth\textsuperscript{a}, Heye Zhang\textsuperscript{b,}\textsuperscript{*}, Shuo Li\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}Western University, London ON, Canada
\textsuperscript{b}Sun Yat-Sen University, Shenzhen, China

Abstract

Accurate and simultaneous segmentation and full quantification (all indices are required in a clinical assessment) of the myocardial infarction (MI) area are crucial for early diagnosis and surgical planning. Current clinical methods remain subject to potential high-risk, nonreproducibility and time-consumption issues. In this study, a deep spatiotemporal adversarial network (DSTGAN) is proposed as a contrast-free, stable and automatic clinical tool to simultaneously segment and quantify MIs directly from the cine MR image. The DSTGAN is implemented using a conditional generative model, which conditions the distributions of the objective cine MR image to directly optimize the generalized error of the mapping between the input and the output. The method consists of the following: 1) A multi-level and multi-scale spatiotemporal variation encoder learns a coarse to fine hierarchical feature to effectively encode the MI-specific morphological and kinematic abnormality structures, which vary for different spatial locations and time periods. 2) The top-down and cross-task generators learn the shared representations between segmentation and quantification to use the commonalities and differences between the two related tasks and enhance the generator preference. 3) Three inter-/intra-tasks to label the relatedness discriminators are iteratively imposed on the encoder and generator to detect and correct the inconsistencies in the label relatedness between and within tasks via adversarial learning. Our proposed method yields a pixel classification accuracy of 96.98\%, and the mean

\textsuperscript{*}Corresponding author

Email address: slishuo@gmail.com (Shuo Li)
absolute error of the MI centroid is 0.96 mm from 165 clinical subjects. These results indicate the potential of our proposed method in aiding standardized MI assessments.

**Keywords:** Myocardial infarction, Segmentation, Full quantification, Sequential images, Generative adversarial networks

1. **Introduction**

Accurate and simultaneous segmentation and full quantification of myocardial infarction (MI) areas (all indices are required in a clinical assessment, including infarct size, percentage of infarct size, percentage of segments, perimeter, centroid, major axis length, minor axis length, orientation and transmurality (Ørn et al., 2007)) are the most effective tools to help clinicians quantitatively and qualitatively assess MI areas (Karim et al., 2013; Heiberg et al., 2008; Schuijf et al., 2004). MI frequently leads to fatal heart failure, and the success of its therapy depends highly on the potential assessment of the scar status. In recent years, clinical studies have indicated that the accurate segmentation of MI areas is important because it reveals the infarcted tissues and facilitates the identification and recovery of the dysfunctional but still viable myocardium (Karim et al., 2013; Ingkanisorn et al., 2004). However, segmentation alone cannot provide all necessary scar information for further treatment. The accurate estimation of all indices of the MI area has gained increasing attention because it reveals the presence, location, and transmurality of acute and chronic MI to improve the selection of the appropriate therapeutic options (ablation or resynchronization) (Heiberg et al., 2008; Flett et al., 2011). Thus, both segmentation and full quantification of MIs are indispensable for early MI patient management and therapy planning because they provide all required information for a thorough understanding of MIs in clinical studies.

Current clinical procedures for MI patients relying on late gadolinium enhancement (LGE) cardiac magnetic resonance (MR) imaging, which is a ‘gold standard’ approach to MI area imaging, are subject to potential high-risk, non-repeatable and time-consuming problems (Fox et al., 2010; Kali et al., 2014; Ordovas and Higgins, 2011), as shown in Fig. 1. A potential high risk caused by the intravenous administration of gadolinium-based contrast agents during the imaging of MI patients (patients after the initial diagnosis by an electrocardiograph) is that this administration can be fatal to patients with chronic kidney diseases (Fox et al., 2010). Patients with coronary artery disease have a higher prevalence of coexisting chronic kidney disease than patients without coronary artery dis-
Figure 1: The proposed method can segment non-contrast agents and quantify an infarction directly from a cine MR image. The advantages of this approach are the great benefit to patients with MI compared to existing LGE-based clinical procedures.

The process is non-repeatable because it requires an experienced clinician to manually/semi-automatically segment the MI areas from the LGE MR image, and this process is subject to high inter-observer variability and subjectivity (Ondovas and Higgins, 2011). The process is also time consuming because it involves three steps: LGE imaging, manual/semi-automatic segmentation from the LEG, and manual/semi-automatic quantification from the segmentation itself to obtain the final quantification result, which may even result in an accumulative error (Kali et al., 2014).

To overcome the aforementioned problems of current clinical procedures, several methods have been attempted to directly segment the MI area from cine MR
images. Cine MR imaging is one of the most common myocardial function examinations, and it presents accurate myocardial spatiotemporal changes. Ultimately, these spatiotemporal changes in myocardial viability caused by MI directly affect the electrical and muscle contractions and drive the morphology and kinematic differences between the MI area and the healthy myocardium (Bijnens et al., 2007). The use of cine MR images in successfully detecting and locating infarcted areas without contrast agents has been reported (Suinesiaputra et al., 2017; Wong et al., 2016). However, this approach has not been sufficiently exploited. Only two cine MR-based methods obtained segmentation of the MI with comparable accuracy to LEG MR (Duchateau et al., 2016; Xue et al., 2017), and no previous method enables one to simultaneously accurately segment and quantify infarcted areas simultaneously.

Accurate and simultaneous segmentation and full quantification of infarcted areas in MI patients directly from cine MR images is extremely challenging and involves the following issues: 1) Systematic learning of the spatiotemporal features of the MI. There were significant differences and inconsistencies in the spatial and temporal correlation of MI in different cine MR images. Since there is no a priori information of the infarcted area in the cine MR images (Petitjean and Dacher, 2011), the effective learning of the hidden spatiotemporal information of cine MR images becomes a prerequisite to establish the intrinsic representation of the MI to differentiate infarcted areas from other tissues. 2) Direct integration of the contiguous feature map of different outputs. There was great heterogeneity and locally optimal discrepancy while training on the input 2D+T (two-dimensional + time) cine MR mapping two different outputs (1D and 2D) (Luc et al., 2016). Accurate learning of the mapping between the segmented image or estimated indices and the cine MR images highly depends on the ability to leverage the adjacency in the output label maps to guide the optimization of feature maps in the training process, such as the optimization of fine details in the segmentation feature maps. 3) Effective exploitation of the significant commonalities and differences between segmentation and quantification tasks (Xue et al., 2018). Segmentation and quantification as two related tasks that share commonalities can use a shared representation during the learning process to produce a beneficial interaction and improve the learning efficiency and prediction accuracy.

To address these challenges, a deep spatiotemporal generative adversarial network (DSTGAN) is proposed to jointly segment and quantify MIs directly from cine MR images without the use of contrast agents. The DSTGAN investigates the infarcted area segmentation/quantification problem from the perspective of conditional generative adversarial networks (GANs), which learn a conditional
Figure 2: The proposed DSTGAN is a conditional generative model. It will be flexible enough to simultaneously segment the image and quantify the index by conditioning on both the generator and three discriminators.

generative model (Fig. 2) (Mirza and Osindero, 2014). The reason for using this model is that the rapid development of GANs has shown impressive results in continuous/discrete data generation (Goodfellow et al., 2014), particularly for assessing the joint configuration of multi-label variables and enforcing forms of higher-order consistency (Luc et al., 2016). As shown in Figure 1, for a query cine MR image of any patient who is preliminarily diagnosed with MI, the DSTGANs will be flexible enough to simultaneously obtain a segmented image and quantified indices by conditioning the spatiotemporal representation of the MI on the input and correspondingly generated output. Specifically, the cine MR images are first mapped to a latent vector through an encoder; then, the vector is projected to segmentation and quantification through a deconvolutional/regression generator. Three adversarial networks are imposed on the generator, which force the generated results to be closer to clinical standards and the distribution of input cine MR images. The contribution of the proposed DSTGANs can be summarized from four aspects:

- For the first time, a non-contrast agent MI simultaneous segmentation and quantification approach is proposed to provide clinicians with all images or indices required in the current clinical assessment of MI directly from 2D+T cine MR images.

- An inherent hierarchical and multi-scale 2D+T encoder is proposed for the
learning of a strong and comprehensive spatiotemporal pyramid representation of 2D+T images. It systematically models the time-series image sequence with large rotation, distortion and different local correlation structures between consecutive frames in different spatial locations and time periods.

- A reciprocal cross-task architecture is proposed to improve the learning efficiency and generation accuracy by exploiting the commonalities and differences across tasks. It merges the deep and coarse information with the shallow and fine information of the MI throughout the network to enrich this beneficial interaction representation.

- An iterative conditional adversarial learning strategy is proposed to integrate three different target discriminators to flexibly use multiple inductive biases to simultaneously and directly approximate the generated distributions to the clinical standards.

This work advances our preliminary work in MICCAI-2018: 1) A new powerful spatiotemporal feature learning network now encodes cine MR images and improves the segmentation/quantification. 2) Conditional GANs have been used in the adversarial training process to enhance the generated results. 3) Three iterative discriminators optimize different distribution tasks to make the model more flexible, while the iterative training strategy ensures that their training is stable. 4) Two additional quantitative indicators (percentage of infarct size and transmurality) have been estimated to cover all indicators that can be used in the clinic. To our knowledge, this is the first report of these two indicators being directly estimated from the cine MR images of MI patients. 5) Experiments are extended to a larger database, and more statistical validations with rigorous discussion have been added to this paper.

2. Related Work

2.1. Existing MI segmentation or quantification methods

Currently, there is no method for simultaneous segmentation and quantification of the MI directly from a cine MR image, and even a direct MI quantification method is not available (Fig. 3). Existing MI segmentation methods cannot satisfy clinical needs in terms of accuracy because these methods provide sparse movement, and the deformation information cannot accurately detect detailed abnormal heart wall motion. These methods mainly fall into three categories: energy-
based (Ledesma-Carbayo et al., 2005), statistical shape model-based (Suinesiaputra et al., 2017), and machine learning-based (Bleton et al., 2015). However, until now, only Duchateau et al. (Duchateau et al., 2016) and Xu et al. (Xue et al., 2017) have successfully achieved pixel-wise MI segmentation through the LV shape uncertainty model and patch-based recurrent neural networks (RNNs), respectively. However, these two methods face the issue of insufficient accuracy because of the cumbersome LV shape modeling process via manual or semi-automated endocardial/epicardial delineation and only single-scale correlation learning for both spatial and temporal features. Moreover, all existing MI quantification methods are based on the quantitative segmentation results (Kali et al., 2014). Compared to this multi-step method, direct quantification in one-step regression is more efficient and more accurate because it avoids the cumulative error generated by multiple steps. Because of the excellent ability of deep learning in task-aware feature representation, direct quantification has also been widely recognized in the medical image community and used in the estimation of different tissues or lesions (Schlegl et al., 2017; Sun et al., 2017). Specifically for the cardiac field, success in using direct quantification has also been reported for estimating a wide range of
indices (Zhen et al., 2015b, 2017a; Afshin et al., 2012), including volume (Zhen et al., 2017b), wall thickness (Xue et al., 2017) and ejection fraction (Gu et al., 2018), from various cardiac images (Zhen et al., 2016, 2017b, 2015a), which led to a challenge called LV full quantification at STACOM-MICCAI’18 (Xue et al., 2017).

2.2. Deep spatiotemporal feature learning

The existing deep spatiotemporal feature learning methods, mainly divided into RNN-based (Yeung et al., 2016) and 3D convolution (Conv)-based (Tran et al., 2015) methods, demonstrate powerful performance (Xingjian et al., 2015; Shou et al., 2016). However, most of these methods cannot be mechanically applied to MI feature learning: 1) The spatial part of spatiotemporal features that are missing or have incomplete results in the RNN-based methods are incapable of modeling the morphological and kinematic changes specific to the MI. The RNN and its variants (such as long short-term memory (LSTM) and convolution RNN (Xingjian et al., 2015)) have been widely used to model the temporal state transitions in frames to capture motion information and improve the inter-frame consistency. However, LSTM fails in taking spatial features into consideration because the fully connected layer learns the 2D+T images as a one-dimensional sequence. Convolution RNN methods, in which the convolution layer is a local constant (fixed kernel size) filter, lack the ability to model the non-local spatiotemporal features specific to the MI with long-range rotation or distortion and different local correlation structures in different spatial locations and time periods (Shou et al., 2016). 2) The loss of granularity over time results in the 3DConv-based model being unable to systematically model all 25 frames from the beginning to the end. 3DConv has been used to capture temporal scene changes when actions are being performed; it is commonly placed into the 3D convolutional neural network framework, which can learn the spatiotemporal abstraction of high-level semantics directly from 2D+T images (Shou et al., 2016). However, 3DConv loses granularity during the temporal growth (Shou et al., 2017).

2.3. Adversarial label relatedness reinforcement

GANs have achieved much success in reinforcing and correcting inconsistent label relatedness in training feature maps. Although the conditional random field (CRF)-based methods also enable reinforcing label relatedness and demonstrate impressive performance when labels are independently predicted, CRF-based methods have limitations in our application because the learned parameters of higher-order potentials are limited in number when both segmented and
quantified label relatedness are integrated into 2D+T data. Thus, the use of an adversarial training approach enforces the multi-range label relatedness without being limited to a notably specific class of high-order potentials in a CRF model, which is a wise choice. The GAN work introduces a new framework for estimating generative models via an adversarial process (Goodfellow et al., 2014). It is composed of two models that are alternatively trained to compete with each other. The generator $G$ captures the distribution of training samples and learns to generate new samples to imitate the training. The discriminative model $D$ differentiates the generated samples from the training. $D$ is optimized to distinguish training samples and to generate results generated by $G$ using a two-player min-max game with the following objective function:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$

(1)

where training data are represented by $x$, the data distribution is $p_{data}$, and $z$ is a vector sampled from a certain distribution $p(z)$.

In conditional GANs (Mirza and Osindero, 2014), by conditioning the model on additional information, it is possible to direct the data generation process. Conditional GANs successfully solve the instability of the original GAN, for which the training process is unstable and the generated results are often noisy and incomprehensible.

3. Methodology

The structure of the DSTGANs is shown in Fig. 4. The input of DSTGANs is a 2D+T cine MR dataset $X$, and the output is $Y = \{y_1, y_2, \ldots, y_n\}$, which includes a segmentation task $y_s = y_1$ and a quantification task $y_q = y_2, \ldots, y_n$, n=10 (including infarct size, percentage of infarct size, percentage of segments, perimeter, centroid, major axis length, minor axis length, orientation and transmurality). DSTGANs apply the GAN model in the conditional setting via three different functional parts: 1) The spatiotemporal variation encoder (STVE) $E(.)$ simultaneously encodes the enriched local correlation structure of different spatial locations and temporal periods in $X$ through spatiotemporal variation learning with a pyramidal structure (Goodfellow et al., 2016). The output of STVE $z$ preserves a comprehensive representation of input $X$, which includes all feature maps from the low level to the high level. 2) The cross-task generator (CTG) $G(.)$ generates the segmentation task and quantification task results simultaneously and
Figure 4: The proposed DSTGAN effectively integrates the spatiotemporal variation encoder, cross-task generator, and three intra-/inter-task label relatedness discriminators to accurately segment (seg) and quantify (Quan) MIs.

reciprocally from the pyramid feature map $z$ via a top-down pathway, and multiple connections upsample and connect the beneficial interaction feature maps of these two related tasks. The output of the CTG can be expressed as $G(z) = Y$.

3) Three discriminators (discriminator segmentation $D_s$, discriminator quantification $D_q$, discriminator relationship $D_r$) are conditioned on $X$ for iterative imposition on $G$. Regarding all of them, $D_s$ and $D_q$ force regularization of the label relatedness within segmentation $y_1$ and quantification $y_2, ..., y_n$ to conform to the relatedness in ground truth, respectively; $D_r$ regularizes the relationship between $y_1$ and $y_2, ..., y_n$ to conform to those in the ground truth.

3.1. Spatiotemporal variation encoder (STVE) to enrich the feature extraction of MI

The STVE innovatively leverages 4 spatiotemporal variation convolution blocks (STVCB) to build a pyramidal feature to accurately learn the left ventricular morphology and kinematic abnormalities. As shown in Fig. 5, this framework with four STVCBs uses a spatiotemporal scaling step of 2 to sequentially encode the cine MR images $X$ from high-resolution to low-resolution and outputs the four spatiotemporal resolution feature maps respectively. Then, these four feature maps combine into a set of pyramidal features $z$. Specifically, the four STVCBs are block1, block2, block3, and block4 and correspond to the output four spa-
Figure 5: The proposed STVE systematically models different local correlation structures of different spatial locations and time periods of MI and other tissues in successive frames using the innovative design of multiple spatiotemporal variation convolution blocks and the spatiotemporal feature pyramid structure. The STVE maps the input cine MR images to a feature pyramid Z.

Spatiotemporal feature maps \( \{C_1, C_2, C_3, C_4\} \), where the spatiotemporal resolution of those pyramid features are \( \{32 \times 32 \times 16, 16 \times 16 \times 6, 8 \times 8 \times 3, 4 \times 4 \times 1\} \) respectively.

The benefit of such pyramid features is two-fold: First, by using multiple scale feature learning frameworks, the encoder can handle the asymmetry distortion (Falsetti et al., 1981), which is a mutability myocardial motion and deformation pattern caused by MI. Second, by using strided convolution instead of the pooling layers, the encoder can be fully differentiable and learn its own spatial downsampling.

Each STVCB integrates multiple spatiotemporal scale sliding sub-blocks that consist of 3DConvs and aConvLSTM to produce different kernel sizes and avoid defects when they are separately used. Multiple sliding convolutional windows of different kernel sizes are used for multiple associations to learn the feature map from the input of a block. Different sliding windows with different kernel sizes proportionally scale the spatiotemporal resolution of the STVCB to simultaneously learn the spatiotemporal dependency structure with different spatial scales.
and multiple time periods. Specifically, 3 scales and 3 aspect ratios are the default for each STVCB. Since \( W_0, H_0, \) and \( T_0 \) are the inputs of the block, the scale ratios are \( \left[ \frac{1}{2}, \frac{1}{4}, \frac{1}{4} \right] \) for all \( W_0H_0T_0 \), and the aspect ratios are \( (1, 1, 2), (1, 2, 1), (2, 1, 1) \) for both scale ratios \( \left[ \frac{1}{2}, \frac{1}{4} \right] \). ConvLSTM is selected for convolutional windows with a scale ratio of \( \{1\} \), which is the input feature map of an STVCB. This selection is natural because ConvLSTM can handle the long temporal information with the spatial correlation better than 3DConv. Simultaneously, the other scales and aspect ratios are used for 3DConv because it can be directly systematically modeled for a short time period by simply adjusting the size of the convolution kernel.

3.2. Cross-task generator (CTG) to share the commonalities between the tasks

The CTG top-down upsamples and merges feature maps from the spatiotemporal feature pyramid \( z \), and it reciprocally connects the features of different layers and different tasks to jointly generate the MI segmentation and quantification results, as shown in Fig. 6.

This CTG efficiently integrates coarse, high-layer information with fine, low-layer information to avoid contextual information loss and gradient dispersion.

- For the segmentation task, the component of CTG is similar to the U-Net (Ronneberger et al., 2015). It begins in a coarser-resolution feature map \( C4 \) and merges rest of the \( z(C3, C2, C1) \) to the same spatial size features during the iterated upsampling processing, where the upsampling size is 2 times. In other words, the CTG merges the feature map of each upsampling unit output with the corresponding \( \{C_1, C_2, C_3, C_4\} \) maps (a 1 × 1 convolutional layer to unify dimensions) and feeds merged features into the next upsampling unit.

- For the quantification task, the component of CTG is modified from the FPN (Lin et al., 2017). It uses five fully connected layers to extract five quantification-related feature vectors \( \{Q_1, Q_2, Q_3, Q_4, Q_5\} \) from five scales of feature maps that are \( C4 \) and four outputs of the upsampling unit. In addition, it then combines these five features to regress the final quantification result by lateral connections. All dimensionalities of the fully connected layers are 256.

Moreover, the CTG also exploits the detailed commonalities and differences between the segmentation task and quantification task to produce the beneficial
interactions among the tasks by shared training. Both segmentation and quantification components of CTG use shared upsampling output and the pyramid feature $z$, and the weights and bias of these two related tasks forcibly share and mutually improve during training.

3.3. Iterative label relatedness discriminators to regularize the contiguous generation of intra-inter-tasks

Iterative label relatedness discriminators ($D_s$, $D_q$, and $D_r$) leverage the intrinsic label relatedness between and within the segmentation and quantification to regularize the generated results (Fig. 7), which effectively improves the accuracy and generalization of the generation. The three discriminators model three types of intrinsic label relatedness: 1) $D_s$ models intra-task relatedness within pixels of the segmentation image (such as spatial structure information); 2) $D_q$ models intra-task relatedness within values of quantification indices (such as the geometric sequence between the area and perimeter); 3) $D_r$ models inter-task relatedness between segmentation and quantification (such as the relationship between the segmentation perimeter and quantification perimeter). The three discriminators check that these relatedness characteristics are consistent with the ground truth, and then improve the generation performance by improving these relatedness characteristics of the generated results until they match the ground truth during
adversarial training. All discriminators condition the spatiotemporal distribution of the input cine MR dataset (after 5 ConvLSTM-Batch Norm-ReLU).

The discriminators $D_s$ and $D_q$ use the CNN and Bi-LSTM (Huang et al., 2015) to build the intra-task label relatedness for segmentation and quantification, respectively. For $D_s$, since the CNN-based model (4 layers, 3 × 3 kernel, max-pooling, stride 2) can assess the spatial correlation of label variables in the field of view, which is the entire image or a large portion of the image, mismatches in the label relatedness statistics can be penalized by the adversarial loss term, which is the binary cross-entropy. For $D_q$, since the Bi-LSTMs (3 layers) can consider a complete contextual relationship to learn a vector pattern, the adversarial process forces the distribution of the generated $y_q$ to gradually approach the prior distribution of ground truth. Note that $c$ and $y_q$ are uniformly scaled to a uniform dimension $d = 100$ via duplicate or 1DConv.

Discriminator $D_r$ uses the Bi-LSTM to build the inter-task label relatedness between segmentation and quantification. Specifically, the segmentation results $y_s$ are fed into two downsampling Bi-LSTM units with dimension $d$. Then, these features are concatenated along the channel dimension with the scaled quantification features and fed into three layers of Bi-LSTMs to jointly learn the features across the image and the indices. Finally, a fully connected layer with one node is used to produce the decision score. Such an inter-task label relatedness learning network can help the CTG rectify small defects in the generated results, which enhances more details in the MI boundary areas and makes the quantification more stable.

3.4. Task-aware loss function to effectively and steadily promote the network training

Our task-aware loss function (Eq. 7) is dedicatedly formulated to train DST-GAN stability and accuracy by combining three loss terms, which are multi-task generation loss (Eq. 2), adversarial loss (Eqs. 3 and 4), and reciprocal loss (Eq. 5).

Generation loss: To improve the accuracy of the segmented images and quantified indices, the multi-task generation leverages the elastic net loss function that integrates L1 (lasso) and L2 (ridge) regularizations to obtain a sparse model that better encourages less blurring and improves the regression performance when it predicts the related tasks.
Figure 7: Three inter/intra-task label relatedness discriminators effectively improve the performance of the generator by directly regulating the distribution of results to approximate the clinical standards.

The generation loss $\mathcal{L}_g(E, G, X)$ are formulated as follows:

$$\mathcal{L}_g = \min_{E, G} \langle \tilde{Y}, G(E(X)) \rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\|\tilde{y}_i - y_i\|_1 + \|\tilde{y}_i - y_i\|_2^2)$$

(2)

where $\tilde{Y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)$ is the ground truth.

Adversarial loss. To train the adversarial training more stably, the adversarial loss is adopted from WGAN-GP (Gulrajani et al., 2017) for three discriminators. In particular, WGAN-GP avoids potential discontinuity with respect to the generators parameters and local saturation leading to vanishing gradients by using a continuous Earth-Mover distance to replace the Jensen-Shannon divergence distance in the GAN formulation. It also adds a gradient penalty to penalize the norm of the gradient of the critic about its input, thereby avoiding undesired behaviors. Thus, adversarial loss $\mathcal{L}_a$ of inter-task label relatedness $\mathcal{L}_{a1}(E, G, X, D_s, D_q)$ and
intra-task label relatedness $L_{a2}(G, X, Y, D_r)$ formula is given as:

$$
L_{a1} = \mathbb{E}_{E(X) \sim P(X)} \left[ D_{\{s,q\}}(G(E(X))) \right] - \mathbb{E}_{\hat{y}_{(s,q)} \sim P_{\{s,q\}}} \left[ D_{\{s,q\}}(y_{(s,q)}, X) \right] + \lambda_1 \mathbb{E}_{\hat{y}_{(s,q)} \sim P_{\hat{y}}} \left[ \left( \| \nabla_{\hat{y}_{(s,q)}} D_{\{s,q\}}(\hat{Y}_{(s,q)}) \|_2 - 1 \right)^2 \right]
$$

(3)

The $L_{a1}$ is a min-max objective function to train generator $G$ and the discriminator segmentation $D_s$ and the discriminator quantification $D_q$ with condition $X$. The $\hat{I}$ is a penalty on the gradient norm for random samples, and $\lambda$ is a penalty coefficient. The encoded pyramid feature $E(X)$ is imposed on $D_s$ or $D_q$, the distribution of the training data is $P_{\{s,q\}}$, and the distribution of the cine MR is $P(X)$.

Similarly, the discriminator $D_r$, and $G$ with condition $X$ can be trained by:

$$
L_{a2} = \mathbb{E}_{E(X) \sim P(X)} \left[ D_r(G(E(X))) \right] - \mathbb{E}_{\hat{y} \sim P_{\hat{y}}} \left[ D_r(y, X) \right] + \lambda_1 \mathbb{E}_{\hat{y} \sim P_{\hat{y}}} \left[ \left( \| \nabla_{\hat{y}} D_r(\hat{Y}) \|_2 - 1 \right)^2 \right]
$$

(4)

$L_{a2}$ drives the discriminator $D_r$ on both the segmentation image and quantification indices to force the generator to further improve the performance. The $L_{a1}$ of $D_s$, $D_q$ and $L_{a2}$ of $D_r$ use iterative (separate) strategy training.

**Reciprocal loss.** To further improve generalization and stability after adversarial training, the reciprocal loss is used to minimize the inter-task label relatedness difference between the $y_s$ and $y_q$ while maintaining the performance. The intuition behind mutual loss is to maintain the accuracy of the results while reducing the difference between two quantification results: 1) the result directly from network regression and 2) the result computed from the segmentation image. These two sets of results should be the same because they are obtained from same LGE image. Thus, the reciprocal loss as an auxiliary term is added behind the generator loss and adversarial loss. It drives the network to have a reciprocal constraint for these two results and an inductive bias for further improving generalization and stability after obtaining accurate results of adversarial training. Note that the percentage of infarct size, the percentage of segments, and the transmurality are
not included. The reciprocal loss $\mathcal{L}_r(G, X, y_q, y_s)$ is defined as:

$$\mathcal{L}_r = \mathbb{E}_{\tilde{y}_q \sim P_{\tilde{Y}}} [||\tilde{y}_q - y_q||_1^2 + ||\tilde{y}_q - y_s||_2^2]$$  \hspace{1cm} (5)$$

where $\tilde{y}_q$ is the quantification results obtained from segmentation image.

**Finally,** To generate the output $Y$, the full objective $\mathcal{L}$ is built to linearly combine all previous partial losses (generation loss $\mathcal{L}_g$, adversarial loss $\mathcal{L}_{a(1,2)}$, and reciprocal loss $\mathcal{L}_r$):

$$\mathcal{L} = \lambda_2 \mathcal{L}_g(E, G) + \lambda_3 \mathcal{L}_{a(1,2)}(E, G, X, Y, D_x, D_q, D_r) + \lambda_4 \mathcal{L}_r(G, X, y_q, y_s)$$  \hspace{1cm} (6)$$

where $\lambda_2$, $\lambda_3$, and $\lambda_4$ are the hyper-parameters that balance the relative importance of the different terms. These hyper-parameters tune the values that achieved the overall highest performance in the experiment. Then, we define the following minimax problem:

$$\arg \min_{E, G} \max_{D \in \{D_x, D_q, D_r\}} \mathcal{L}$$  \hspace{1cm} (7)$$

4. Materials and Implementation Details

**Materials** A total of 165 patients were retrospectively selected from two clinical databases for a retrospective observational study, including 140 MI patients and 25 NON-MI patients. All patients were admitted with acute ST-segment elevation MI (in at least 2 contiguous electrocardiogram leads). All patients completed cine (12375 images) and LGE (495 images) MR imaging scans. Cine MR imaging was performed using a 3T MR system (MAGNETOM Verio, Siemens Healthcare) with a 32-channel cardiac coil. After the scout MR images were obtained, cine images were acquired during repeated breath-holds to cover the entire heart. LGE MR imaging was performed in the same orientations and with the same slice thickness as cine imaging using a segmented inversion-recovery gradient-echo sequence ten minutes after the intravenous injection of gadolinium (Magnevit, 0.2 m mol/kg; Bayer Schering Healthcare). The inversion time was optimized to null the myocardium during the image acquisition. The imaging parameters were as follows: TR = 10.5 ms, TE = 5.4 ms, FA = 30°, matrix of 256×162, slice thickness of 6 mm, BW of 140 Hz/px, and 25 cardiac phases, each pixel is 1.26x1.26 mm.

The ground truth was obtained by two radiologists. In detail, two experienced (over 10 years) radiologists (Dr. N. Zhang and Dr. L. Xu) manually assessed
LGE MR images (the well-known gold standard in the clinical MI assessment) and segment the infarction area to obtain the ground truth of segmentation, and if a disagreement occurred, it was resolved by a consensus between the experts. Then, Dr. N. Zhang measured the ground truth of the segmentation to obtain the ground truth of the quantification.

**Implementation details.** DSTGANs are designed to generate $64 \times 64$ binary images and $1 \times 10$ indices directly from $\mathbf{X} \in (x_1, x_2, \ldots, x_T, \mathbb{R}^{H \times W \times T})$, where $H$ and $W$ are the height and width of each temporal frame, respectively ($H = W = 64$), and $T$ is a temporal step set to $T = 25$. The input vector is first encoded to a feature of the pyramid, which consists of a $32 \times 32 \times 256T$, $16 \times 16 \times 256T$, $8 \times 8 \times 256T$, $4 \times 4 \times 256T$ tensor, where $T$ is the number of temporal resolutions in the tensor. Batch normalization (Goodfellow et al., 2016) and LeakyReLU activation (Goodfellow et al., 2016) are applied after every block of the encoder except the last one. The residual blocks consist of 33 stride 1 convolutions, batch normalization and LeakyReLU. The upsampling blocks consist of nearest-neighbor upsampling followed by a 33 stride 1 convolution. All codes are based on Keras (Tensorflow), with an 8x NVIDIA P100 for training. All networks are trained using the ADAM solver (Kingma and Ba, 2014) with a batch size of 6 and an initial learning rate of 0.001. The $\lambda_1$ is 10, $\lambda_2$ is 0.5, $\lambda_3$ is 1, and $\lambda_4$ is 10. The iterative training strategy is used, and stage 1 iteratively trains $\{E, G\}$ and $\{D_s, D_q\}$ for 800 epochs by fixing $D_r$. Then, stage 2 iteratively trains $G$ and $D_r$ for another 600 epochs by fixing $\{E, G\}$ and $\{D_s, D_q\}$. There are two iterative training processes. A network with heart localization layers, as described in (Xu et al., 2017), was used to automatically crop the cine MR images to $64 \times 64$ region-of-interest sequences including the LV. Moreover, because LGE MR imaging was performed in the same orientations and with the same slice thickness as cine imaging, no extra steps were required to align them. All experiments were assessed with a 10-fold cross-validation test (no patients are in different groups).

**Evaluation criteria.** The performance of the proposed DSTGAN framework is evaluated based on the generation accuracy of the MI segmentation and quantification. For the segmentation task, the overall pixel classification accuracy, sensitivity, specificity, Dice coefficient ($\text{Dice} = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \times 100\%$) and intersection-over-union ($\text{IoU} = \frac{|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \times 100\%$) between the ground truth and the generated image are calculated to evaluate the accuracy. For the quantification task, the mean absolute error ($\text{MAE} = \frac{\sum_{n=1}^{n}|Y - \hat{Y}|}{n}$) and statistical hypothesis testing ($p$-value) between the ground true and generated indices are calculated to evaluate
the accuracy. Note that the MAE of the centroid is the sum of the MAE in X-axis and the MAE in Y-axis. The $p$-value is obtained by an independent t-test between the prediction values and the ground truth. If it is greater than 0.05 (i.e., $p > .05$), the two groups can be treated as consistent. If $p < 0.05$, the two groups violate the assumption of homogeneity of variances and can be treated as inconsistent.

5. Experiments and Results

The DSTGAN demonstrates high performance with a pixel classification accuracy of 96.98%. The MAE of the centroid point is 0.96 mm with the ground truth obtained manually by human experts, which demonstrates the effectiveness of this method in MI segmentation and quantification.

5.1. Accurate MI segmentation and precise MI quantification

The experimental result shows that the DSTGAN can accurately segment the MI, as shown in Figure 8. The DSTGAN achieves an overall pixel classification accuracy of 96.98% with a sensitivity of 92.15% and a specificity of 98.70%. The Dice coefficient is 92.89±0.33%, and the IoU is 82.16±0.27%; the ROC and PR curves are shown in Fig. 9. The ground truth and the result are binary images; each pixel is assessed for infarction or normality (0 or 1).
DSTGANs can produce a good quantification of the MI, as shown in Tables 1 and 2. The MAE computed between the ground truth and our estimation of the infarction size is 17.15±8.91 mm²; the percentage of the infarcted size is 9.24±7.73%; the percentage of the segment is 4.07±2.91%; the perimeter is 3.95±2.21 mm; the centroid is 0.96±0.77 mm; the major axis length is 2.04±1.10 mm; the minor axis length is 1.07±0.80 mm; the orientation is 5.47±2.24°; and the transmurality is 13.18±9.83%. Table 3 summarizes the comparative results on each quantification index subgroup. The ground truth was obtained by radiologists who measured the segmentation results of DE-MR images.

Tables 1 and 2 also demonstrated that the joint learning of the segmentation task and quantification task outperforms each individual task. The joint learning improves the effectiveness and stability for both segmentation and quantification in the terms of accuracy (1% improved), Dice (2% improved), IOU (3% improved), infarcted size (1 mm improved), the percentage of the infarcted size (4% improved), the percentage of the segment (0.5% improved), the perimeter (1 mm improved), the major axis length (0.1 mm improved), the minor axis length (0.7 mm improved), the orientation (1° improved), and the transmurality (3% improved).
Table 1: Each technological innovation component in DSTGANs has effectively improved the accuracy of the MI segmentation and part of the quantification

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Accuracy (%)</th>
<th>Dice (%)</th>
<th>IOU (%)</th>
<th>Size (mm²)</th>
<th>Per-size (%)</th>
<th>Per-Seg (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvLSTM (Baseline)</td>
<td>89.27</td>
<td>83.70</td>
<td>66.10</td>
<td>72.73</td>
<td>10.54</td>
<td>8.91</td>
</tr>
<tr>
<td>+ STVCB</td>
<td>90.94</td>
<td>83.68</td>
<td>69.63</td>
<td>80.37</td>
<td>11.63</td>
<td>6.95</td>
</tr>
<tr>
<td>+ Feature pyramid structure</td>
<td>92.37</td>
<td>85.68</td>
<td>70.51</td>
<td>49.32</td>
<td>10.14</td>
<td>4.46</td>
</tr>
<tr>
<td>+ CTG</td>
<td>95.10</td>
<td>90.68</td>
<td>73.47</td>
<td>27.26</td>
<td>12.64</td>
<td>5.17</td>
</tr>
<tr>
<td>+ Dₐ</td>
<td>95.92</td>
<td>91.26</td>
<td>77.34</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+ D₉</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>18.72</td>
<td>10.11</td>
<td>4.62</td>
</tr>
<tr>
<td>+ Dₐ and D₉</td>
<td>96.71</td>
<td>91.43</td>
<td>79.52</td>
<td>19.24</td>
<td><strong>8.76</strong></td>
<td>4.34</td>
</tr>
<tr>
<td>+ Dₐ</td>
<td>96.93</td>
<td>92.11</td>
<td>78.85</td>
<td>20.61</td>
<td>9.69</td>
<td>1.42</td>
</tr>
<tr>
<td>+ Conditional</td>
<td>96.98</td>
<td><strong>92.89</strong></td>
<td><strong>80.16</strong></td>
<td><strong>17.15</strong></td>
<td><strong>9.24</strong></td>
<td><strong>4.07</strong></td>
</tr>
</tbody>
</table>

MAE is used to evaluate all the quantification indices
Size = infarct size. Per-size = percentage of infarct size. Per-Seg = percentage of segments

5.2. Advantages of the spatiotemporal feature encoder

Figures 9 and 10 and Tables 1, 2 and 3 indicate that the STVE performs better than all other frameworks because it creates a pyramid structure to produce a coarse to fine hierarchical feature and achieves a strong spatial and temporal representation of multiple scales and aspect ratios. To evaluate the feature-learning performance, we replace the STVE with a version without the pyramid structure and without STVCBs and with recently popular methods (3DConvs, ConvLSTMs, 3DConvs + LSTMs, CNNs and LSTMs) in our framework. These experimental results demonstrate that the integration of different spatiotemporal scales and aspect ratios and the hierarchical, coarse-to-fine spatiotemporal feature learning
Table 2: Table 1 Continued

<table>
<thead>
<tr>
<th>Quantification</th>
<th>Peri (mm)</th>
<th>Cent (mm)</th>
<th>Majo (mm)</th>
<th>Mino (mm)</th>
<th>Ori (°)</th>
<th>Tran (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvLSTM (Baseline)</td>
<td>10.17</td>
<td>4.55</td>
<td>9.49</td>
<td>8.64</td>
<td>12.62</td>
<td>18.61</td>
</tr>
<tr>
<td>+ STVCB</td>
<td>±8.20</td>
<td>±2.97</td>
<td>±7.31</td>
<td>±7.08</td>
<td>±10.35</td>
<td>±15.96</td>
</tr>
<tr>
<td>+ Feature pyramid structure</td>
<td>7.37</td>
<td>2.78</td>
<td>6.84</td>
<td>5.35</td>
<td>8.27</td>
<td>15.07</td>
</tr>
<tr>
<td>+ CTG</td>
<td>±6.94</td>
<td>±1.28</td>
<td>±5.01</td>
<td>±3.50</td>
<td>±6.22</td>
<td>±13.72</td>
</tr>
<tr>
<td>+ $D_s$</td>
<td>±5.40</td>
<td>±0.72</td>
<td>±5.49</td>
<td>±3.38</td>
<td>±7.15</td>
<td>±16.83</td>
</tr>
<tr>
<td>+ $D_q$</td>
<td>8.61</td>
<td>1.13</td>
<td>5.34</td>
<td>2.23</td>
<td>7.74</td>
<td>15.27</td>
</tr>
<tr>
<td>+ $D_s$ and $D_q$</td>
<td>±6.68</td>
<td>±0.88</td>
<td>±3.97</td>
<td>±1.39</td>
<td>±5.81</td>
<td>±13.35</td>
</tr>
<tr>
<td>+ Conditional</td>
<td>±2.28</td>
<td>±0.91</td>
<td>±1.50</td>
<td>±0.94</td>
<td>±4.29</td>
<td>±10.92</td>
</tr>
</tbody>
</table>

MAE is used to evaluate all the quantification indices
Peri = perimeter, Cent = centroid, Majo = major axis length, Mino = minor axis length, Ori = orientation, Tran = transmurality

framework is more reliable and robust than the cases that do not use or independently use them. This integration provides our framework with a more comprehensive understanding than all existing methods that learn the morphological and kinematic abnormalities of the left ventricle. Furthermore, our experiments demonstrate that the STVCB, which combines ConvLSTM and 3DConv to simultaneously extract different spatiotemporal resolution features, is the best choice among similar methods. This combination avoids the disadvantages of ConvLSTM with only a single spatiotemporal resolution and 3DConv, which is only suitable for short-term spatiotemporal learning, and directly models the advanced temporal and spatial abstraction of MIs by considering spatial and temporal correlations.
Table 3: STFPNs work better than other popular spatiotemporal learning constructs.

<table>
<thead>
<tr>
<th></th>
<th>Dice</th>
<th>Infacted size (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STVE</td>
<td>92.89%</td>
<td>17.15</td>
</tr>
<tr>
<td>3DConvs</td>
<td>81.67%</td>
<td>65.80</td>
</tr>
<tr>
<td>ConvLSTMs</td>
<td>86.37%</td>
<td>49.78</td>
</tr>
<tr>
<td>3DConvs + LSTMs</td>
<td>87.46%</td>
<td>42.01</td>
</tr>
<tr>
<td>LTMs</td>
<td>74.07%</td>
<td>116.74</td>
</tr>
<tr>
<td>CNNs</td>
<td>71.91%</td>
<td>224.44</td>
</tr>
</tbody>
</table>

5.3. Advantage of the conditional GAN architecture

Figure 9, Tables 1 and 2 show that the DSTGAN has better segmentation and quantification performance than STVE + CTG without the GAN architecture only because it enhances the inter/intralabel relatedness via adversarial learning. To evaluate this ability, discriminators $D_s$, $D_q$ and $D_r$ are individually implemented in the CTG for the estimated tasks. Among these results, Tables 1 and 2 indicate that the integration of the top-down, coarse-to-fine feature map generation process and the cross-task connection sharing process enables improving the accuracy of the task generation. This integration improves the loss of information when a feature map is multiple-sampled or upsampling and improves the shared representation between segmentation and quantification tasks by learning the reciprocal interaction between them. Tables 1 and 2 and Figure 9 indicate that the three types of discriminators effectively improve the accuracy and robustness of the generated results by learning the relatedness within and among the task labels, respectively, and this architecture is flexible and easy to implement. Moreover, Tables 1 and 2 indicate that the conditional information effectively controls the generation process to reduce the fluctuation of the result and improve the robustness of the network.

5.4. Advantage of the jointing using L1 and L2 loss

Table 4 demonstrates that our generator loss of jointly using the L1 and L2 term achieve an overall better solution than only using the L1 term. Jointly using L1 and L2 improves the performance in the aspects of infarction size, the percentage of the segment, the major axis length, the minor axis length, and the orientation. This is because our network involves multitask learning, and L2 of L1 + L2 outperforms L1 in the regression part of the efficiency and effectiveness. Moreover, jointly using L1 and L2 significantly improves the stability for all segmentation and quantification results. This is because the addition of the L2 term
Table 4: Jointly using L1 and L2 loss achieves a better solution than only the L1 term.

<table>
<thead>
<tr>
<th></th>
<th>Dice</th>
<th>Size</th>
<th>Per-size</th>
<th>Per-Seg</th>
<th>Peri</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>92.90 ± 0.51</td>
<td>17.94 ± 9.11</td>
<td>9.06 ± 8.62</td>
<td>4.61 ± 3.27</td>
<td>3.90 ± 2.88</td>
</tr>
<tr>
<td>L1+L2</td>
<td>92.89 ± 0.33</td>
<td>17.15 ± 8.91</td>
<td>9.24 ± 7.73</td>
<td>4.07 ± 2.91</td>
<td>3.95 ± 2.21</td>
</tr>
<tr>
<td></td>
<td>Cent</td>
<td>Majo</td>
<td>Mino</td>
<td>Orie</td>
<td>Tran</td>
</tr>
<tr>
<td>L1</td>
<td>0.95 ± 0.77</td>
<td>2.21 ± 1.11</td>
<td>1.13 ± 0.82</td>
<td>5.49 ± 3.09</td>
<td>13.02 ± 10.52</td>
</tr>
<tr>
<td>L1+L2</td>
<td>0.96 ± 0.77</td>
<td>2.04 ± 1.10</td>
<td>1.07 ± 0.80</td>
<td>5.47 ± 2.24</td>
<td>13.18 ± 9.38</td>
</tr>
</tbody>
</table>

MAE is used to evaluate all the quantification indices
Peri = perimeter, Cent = centroid, Majo = major axis length, Mino = minor axis length, Orie = orientation, Tran = transmutability

Forces the issues that the derivatives of L1 are not continuous in all tasks, thus more stable results achieved.

5.5. Comparison with state-of-the-art methods

Table 5 demonstrates that DSTGANs achieve higher segmentation and quantification accuracy and more quantification terms than the existing methods in segmentation/quantification (Seq/Qua). MuIGAN corresponds to our MICCAI 2018 work. Both terms ‘Seq’ and ‘Qua’ indicate that this method can directly segment or quantify myocardial infarction without additional steps. The term ‘NaN’ indicates that this method cannot estimate this index. Overall, by comparison, DSTGANs consistently outperform our previous method and all state-of-the-art MI segmentation methods (Xue et al., 2017; Popescu et al., 2016; Bleton et al., 2015) in terms of accuracy, Dice coefficient, IoU, infarct size, perimeter and centroid, and the MI’s transmutality and percentage of infarct size of the LV are obtained for the first time. These significant improvements in segmentation, particularly in IoU, indicate that DSTGANs improve the existing methods have large numbers of badly scattered points, unconnected areas and severe overestimation problems (Xue et al., 2017). A possible reason is that our method systematically models the motion and deformation of 2D+T data instead of the pixel-by-pixel estimation in existing methods and uses the inter/intra-label relatedness to regulate the training process. Furthermore, the improvement in MI quantification demonstrates the superiority of DSTGANs as a direct quantification method and a combination of quantification and segmentation. This work exceeds the work in MICCAI-2018 through a more effective spatiotemporal feature learning framework, a more detailed generation structure, and more advanced training strategies.
Table 5: DSGANs simultaneously segment and quantify MIs without a contrast agent and yield higher performance than some state-of-the-art methods.

<table>
<thead>
<tr>
<th></th>
<th>DSGANs</th>
<th>MuTGAN</th>
<th>Xu et al (Xu et al., 2017)</th>
<th>Popescu et al (Popescu et al., 2016)</th>
<th>Bleton et al (Bleton et al., 2015)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seg/Qua Accuracy</td>
<td>96.98%</td>
<td>96.57%</td>
<td>94.73%</td>
<td>86.47%</td>
<td>84.39%</td>
</tr>
<tr>
<td>Dice</td>
<td>92.89%</td>
<td>90.18%</td>
<td>89.40%</td>
<td>75.33%</td>
<td>71.51%</td>
</tr>
<tr>
<td>IoU</td>
<td>82.16%</td>
<td>79.51%</td>
<td>76.28%</td>
<td>68.15%</td>
<td>63.72%</td>
</tr>
<tr>
<td>Infarct size</td>
<td>17.15%</td>
<td>23.24%</td>
<td>93.16*</td>
<td>159.93*</td>
<td>194.75*</td>
</tr>
<tr>
<td>Perimeter</td>
<td>3.95%</td>
<td>5.27%</td>
<td>17.83*</td>
<td>31.62*</td>
<td>44.37*</td>
</tr>
<tr>
<td>Centroid</td>
<td>0.96%</td>
<td>1.16%</td>
<td>6.49*</td>
<td>8.03*</td>
<td>10.92*</td>
</tr>
<tr>
<td>Pct. infarct size</td>
<td>4.07%</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>Transmurality</td>
<td>13.18%</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

* Quantification result estimated from segmentation result, and Pct. = Percentage of
Note that the MuTGAN corresponds to our MICCAI 2018 work.

6. Discussion

To the best of our knowledge, this is the first report that describes a non-contrast agent approach that can provide clinicians with all the required factors for the current clinical assessment of MI directly from cine MR imaging.

Our approach satisfies a clinical desire to replace contrast agent administration with an inexpensive, end-to-end, low-risk, reproducible and steady tool with high clinical impact that can change the clinical procedure and management of MI patients. This is vital for clinical studies because of the limitations of contrast agent administration (e.g., using gadolinium chelate), which may not be suitable for patients with chronic kidney disease according to the US Renal Data System. Specifically, >40% patients with chronic kidney disease also suffer from cardiovascular disease (Fox et al., 2010), and 20% of acute MI patients have accompanying chronic kidney disease (Fox et al., 2010). In contrast, our framework only requires cine images, which are inevitably acquired as a routine part of a cardiac examination for assessment of function, and produces results with high accuracy in segmenting and quantifying MIs that are comparable to the LGE technique. This is evidenced by Table 6, which indicates that no significant differences ($P > 0.05$) were found between the results obtained by our method and the ground truth obtained by experienced experts.

For the first time, our approach provides a clinical tool that can automatically quantify the transmurality and percentage infarct size of the LV of the MI directly from cine MR images. This ability can help clinicians to more accurately 1) plan the patient treatment because these indices are the primary determinant of
Table 6: No significant differences ($P > 0.05$) were found between the quantification results obtained using our methods and the clinician-measured values in the LGE image.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>DSTGANs</th>
<th>Ground Truth</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infarct Size ($cm^2$)</td>
<td>9.31±2.69</td>
<td>7.74±2.20</td>
<td>.146</td>
</tr>
<tr>
<td>Pct. infarct size (%)</td>
<td>28.94±19.31</td>
<td>24.47±16.25</td>
<td>.060</td>
</tr>
<tr>
<td>Pct. segments (%)</td>
<td>27.39±16.43</td>
<td>25.97±12.29</td>
<td>.079</td>
</tr>
<tr>
<td>Perimeter ($mm^2$)</td>
<td>172.55±73.49</td>
<td>170.85±82.79</td>
<td>.102</td>
</tr>
<tr>
<td>Centroid</td>
<td>X:30.24±13.66</td>
<td>X:30.03±13.00</td>
<td>.743</td>
</tr>
<tr>
<td></td>
<td>Y:34.34±10.91</td>
<td>Y:33.72±11.28</td>
<td></td>
</tr>
<tr>
<td>Major axis length ($mm$)</td>
<td>73.74±26.15</td>
<td>72.16±28.42</td>
<td>.515</td>
</tr>
<tr>
<td>Minor axis length ($mm$)</td>
<td>33.64±21.22</td>
<td>31.40±19.99</td>
<td>.260</td>
</tr>
<tr>
<td>Orientation ($^\circ$)</td>
<td>109.44±56.49</td>
<td>106.02±56.28</td>
<td>.112</td>
</tr>
<tr>
<td>Transmurality (%)</td>
<td>77.60±33.54</td>
<td>64.59±27.38</td>
<td>.089</td>
</tr>
</tbody>
</table>

Pct. = Percentage of

LV remodeling in long-term survivors of MI (Örn et al., 2007), 2) diagnose the LV constriction damage because the percentage infarct size of the LV is a major determinant of the LV performance after MI (Örn et al., 2007) and 3) determine the possibility of cardiac functional recovery after revascularization because lower transmurality indicates a high likelihood (Örn et al., 2007). However, the direct quantification of these two indices is extremely challenging for a diagnosis based on nonenhanced imaging because 1) small transmural and subendocardial scars often show no signal abnormalities on cine MR images and 2) additional information on the LV wall that is not included in our training data is required, with one method requiring LV wall thickness and the other requiring LV wall size. Even so, our approach successfully quantifies these two indices directly from cine MR images because it can learn the detailed pixel’s motion and deformation information of all tissues in the cine MR images during training (some of the characteristics of the LV wall have been learned and displayed), as shown in Figure 10.

Our proposed DSTGAN can be extended to similar applications because of the following technical advantages: 1) Multi-scale kernel 3D convolution and spatiotemporal feature pyramids are first constructed to effectively handle the high spatiotemporal variability in different spatial scales and temporal periods, and they overcome the limitation of variability with great gaps in spatiotemporal patterns and disturbances in the pixel gray levels among different locations or periods. 2) The task-aware objective function learns a reciprocal commonality representation of two related tasks to train mappings from the input cine MR images to two re-
Figure 10: The feature map clearly demonstrates the ability of our method to learn the motion and deformation of cine MR images. The second row represents the $32 \times 32$ feature maps of the corresponding segmentation ground truth (first row). The high-intensity regions in these feature maps approximately correspond to the most responsive area for the deformation and motion. The low-intensity regions represent the infarction area because infarction causes deformation and motion abnormalities (motion reduced and myocardium thinned). In addition, our method can also be unsupervised to learn some of the characteristics of the LV wall through the unique spatiotemporal information of the myocardium.

lated outputs and ensures that the generated results indeed approach the ground truth via adversarial training. This function overcomes the limitation of multitasking adversarial training, which requires notably different loss formulations to address different distributions of features, which are difficult to train. 3) Conditional adversarial learning with multiple discriminators successfully iteratively optimizes the integrated distribution with the input as conditional information using the label contiguity to produce deterministic outputs. This learning strategy to avoid a basic conditional GAN has a stochastic output because of noise $z$ and enhances the generation performance.

7. Conclusion

A deep spatiotemporal adversarial network has been proposed for simultaneous segmentation and quantification of MIs without the use of contrast agents. The DSTGAN was conducted on 165 subjects and yielded a pixel classification accuracy of 96.98%; the MAE of the infarction size was 17.15 mm$^2$. These results demonstrate that our proposed method can be an efficient and accurate clinical tool to systemize MI assessment and provide timely interpretations for clinical
follow-up.

References


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.


Conflicts of interest: none