Biomedical Signal Processing and Control 82 (2023) 104594

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Registration of 3D medical images based on unsupervised cooperative cascade of deep networks

Gangcheng Cai[1], Huaying Liu[1], Wei Zou, Nan Hu, JiaJun Wang *

School of Electronic and Information Engineering, Soochow University, Suzhou, PR China

ARTICLE INFO

Keywords:
Unsupervised registration
Convolutional neural networks
Cascades
3D medical image

ABSTRACT

In this paper, a deformable registration network (DR-Net) and a multi-scale cascading strategy are designed for the registration of largely deformed 3D medical images. Our DR-Net appears as a U-shaped convolutional neural network with a pyramidal input module (PIM), a light weighted sequential Inception module and an SCAM convolutional attention module. Our multi-scale cooperative cascading strategy integrates the deformation field information within and between sub-networks at different scales to synthesize the cascaded deformation fields. To cooperatively train the cascaded network, not only the output of the final network layer but also the multi-scale outputs from different layers of the decoder in the last cascaded sub-network are used to calculate loss function. As compared with the VoxelMorph and IVTN, the average dice similarity coefficients (Dice) achieved with our DR-Net are 2.4% and 2.5% higher on the Sliver dataset and are 2.5% and 2.4% higher on the LiTS dataset. The average Dice coefficients achieved with our multi-scale cascading strategy of three DR-Nets are 1.6% and 1.9% higher than those of the VM-CR3 and are 1.5% and 1.7% higher than those of the IVTN-CR3 on these two datasets, respectively. These results show that not only our proposed DR-Net itself but also the cascade of them outperform the state-of-the-art methods and their cascades in registration accuracy.

1. Introduction

Image registration aims to find the non-linear spatial correspondence between fixed and moving images. It has a wide range of applications in medical image processing, such as aligning images of one subject taken at different times. Another example is to match an image of one subject to some predefined coordinate system, such as an anatomical atlas [1]. Accurate and fast registration also plays an important role in guiding medical surgery. However, due to the variable size, shape, and location, as well as the heterogeneity of tissue content of datasets, this technic faces many difficulties. In addition, registration of multimodal medical images is more complex than that of mono-modal images and may face more difficulties [2,3]. Therefore, many researchers and scientists dedicated themselves to finding suitable methods for medical image registration.

Over recent decades, many traditional algorithms have been developed and studied to automatically register medical images [4–7]. Generally, these algorithms define a space of transformations and a metric of alignment quality, and then find the optimal transformation by iteratively updating the parameters. Traditional methods have achieved good performance on several datasets, but their registration speed is rarely suitable for clinical applications especially in 3D cases. During recent years, deep learning based methods have changed the research pattern in the fields of segmentation [8,9], retrieval [10–12] and registration [13,14] of medical images. Deep learning based registration methods learn patterns represented by the parameters of the underlying model to efficiently replace the optimization procedure in traditional methods and hence they run much faster than traditional ones. Nowadays, deep learning is becoming more and more popular in medical image registration. Deep models for medical image registration can be trained in supervised [15–18] or unsupervised manners [13,19–22]. Although supervised learning registration methods have achieved good results, they need ground truth for the deformation field. However, deformation fields are almost impossible to be labeled manually since flow fields are dense and ambiguous quantities. Besides, automatically generated datasets (*e.g.*, the Flying Chairs dataset [17], Flying Things 3D [23]) which deviate from the realistic demands are not appropriate neither. Consequently, supervised methods are hardly applicable. Therefore, more and more researchers dedicated to developing unsupervised methods for medical image registration. To ease the necessity of using the ground truth for the deformation fields, the

* Corresponding author.
E-mail addresses: gccai@stu.suda.edu.cn (G. Cai), 20204228016@stu.suda.edu.cn (H. Liu), zouwei@suda.edu.cn (W. Zou), hunan@suda.edu.cn (N. Hu), jjwang@suda.edu.cn (J. Wang).
[1] These authors contributed to the work equally and should be regarded as co-first authors.

https://doi.org/10.1016/j.bspc.2023.104594
Received 28 August 2022; Received in revised form 18 December 2022; Accepted 9 January 2023
Available online 23 January 2023

unsupervised methods use the discrepancy (similarity loss) between the fixed image and the output warped moving image of the network to supervise the training process. One typical example is that proposed by Shan et al. [19] for deformable registration of 2D MR brain images and CT liver scans which tries to learn sparse parameters introduced by traditional algorithms. VoxelMorph is another unsupervised learning-based method proposed by Balakrishnan et al. [13] for 3D medical image registration which directly uses U-Net [24] to derive the deformation field for warping moving images and is trained to minimize the dissimilarity between the warped moving image and the fixed image. It performs well on brain MRI datasets but badly on liver CT scans especially when large displacement between images is presented. A systematic review of different unsupervised deep registration methods can be found in [25].

Most existing deep networks make a straightforward prediction of the flow field, which makes them rather difficult when handling complicated deformations, especially with large displacements. To solve this problem, de Vos et al. [26] proposed a deep learning based image registration method (DLIR) which stacked deep networks together and was trained one by one (i.e., any cascade was trained by fixing the weights of previous cascades). Although this training scheme is very simple to implement, it only achieves limited performance improvement. Later, the volume tweening network (VTN) [27] tried to train all the networks in the cascade jointly where the similarity between the output of each sub-network and the fixed image was used to generate the training loss to guide the training process of the cascaded networks. However, this training strategy is still a non-cooperative one because each sub-network is trained only under the supervision of the similarity measure between the output of the current sub-network and the fixed image regardless of the existence of other cascaded sub-networks during training stage. Though it was improved in its subsequent version in [28] (Improved Volume Tweening Network, IVTN) where the similarity was only measured on the final warped image to guide the training of the cascaded networks, there still exist some other drawbacks for this cascading strategy: (1) Due to the fact that the final deformation field is synthesized from deformation field of a single scale, it contains insufficient hierarchical deformation information critical for improving the registration accuracy of large deformed images and of local small areas; (2) The training of the whole network is only guided by the loss of the last layer. This results in suboptimal estimation of parameters in the frontal convolution layers since parameters of the first half convolution layers are not updated as much as those of the latter half. This not only deteriorates the registration accuracy but also makes the training process hard to converge; (3) The cascading strategy achieves fairly smaller improvement in registration accuracy at the expense of introducing a larger number of parameters which poses higher hardware demands. With these non-cooperative cascades, further improvement can hardly be achieved even if more sub-networks are incorporated.

To tackle the aforementioned problems especially for registration of images with large deformation, this paper focuses on designing a deformable registration network (DR-Net) and a multi-scale cascading strategy which tries to make the model more trainable for extracting multi-scale features to synthesize deformation field with hierarchical displacement information. Our DR-Net appears as a U-shaped convolutional neural network with a pyramidal input module (PIM), a light weighted sequential Inception module and an SCAM convolutional attention module. The PIM is introduced to enable the network extracting richer and hierarchical image features while the Inception module is used to enable the network integrating features of different scales. With the SCAM, the network can learn to simultaneously highlight useful channels or voxels while suppress irrelevant ones to our registration tasks. Rather than trying to improve the registration accuracy by appending more and more sub-networks in the cascade as usual, we propose a multi-scale cooperative cascading strategy which integrates the deformation field information within and between sub-networks at different scales to synthesize the cascaded deformation fields. To cooperatively train the cascaded network, not only the output of the final network layer but also the multi-scale outputs from different layers of the decoder in the last cascaded sub-network are used to calculate loss function. The main contributions of this paper are as follows:

- We design a light weighted U-shaped deformable registration network (DR-Net) with a pyramidal input module (PIM), a light weighted sequential Inception module and an SCAM convolutional attention module.
- We propose a multi-scale cascading strategy which integrates the deformation field information within and between sub-networks at different scales to synthesize the cascaded deformation fields with more detailed displacement information.
- We propose a cooperative training strategy for the cascaded network where not only the output of the final network layer but also the multi-scale outputs from different layers of the decoder in the last cascaded sub-network are used to calculate loss function.

2. Methodology

Mathematically, the registration of I_m to I_f is to find a mapping (deformation field) $\phi : \Omega \rightarrow \Omega$ so that I_m matches I_f best upon performing warping operations with respect to I_m according to ϕ. In this work, this mapping will be found with deep neural networks.

2.1. Framework of the proposed unsupervised 3D medical image registration strategy

Fig. 1 illustrates the framework of the proposed unsupervised image registration strategy for 3D medical images which contains two stages: the training stage and the evaluating stage. It should be noted that although our experiments are conducted between 3D images, only two-dimensional slices of the images are given in the figure for display convenience. During the training stage, fixed and moving image pairs are sequentially fed into the convolutional neural network for predicting the deformation field under an initial setting of network parameters. Then the predicted deformation field is fed into a spatial transform network (STN) [29] to warp the moving image, after which the registration error between the fixed and warped moving images is calculated and feedback to the network for updating the network parameters in an iterative manner until convergence. In order to avoid overfitting, the regularization loss on the flows are introduced. In the evaluating stage, pairs of images are fed into the trained network to generate the deformation field. The moving images and the corresponding labels will be warped by the obtained deformation field to align with the fixed images and fixed labels. Many metrics, such as Dice Score between fixed labels and warped moving labels, can be employed to evaluate the registration results.

2.2. The architecture of deformable registration network

Both the architecture and the training strategy are key issues affecting the performance of our registration method. The VoxelMorph directly uses the U-Net structure while the IVTN tries to improve the registration accuracy by simply deepening U-Net structure. However, excessively deeper network not only results in excessively smaller-sized feature map that contributes little in improving registration accuracy but also extremely increases the amount of network parameters, which makes the training process difficult to converge. Furthermore, the maximum pooling layers in these two methods lead to the loss of important information. Another issue is that the U-Net type network fuses the low-level and high-level features directly through skip connections while neglecting the semantic disparity between the encoder–decoder features, which significantly limits the performance of the network.

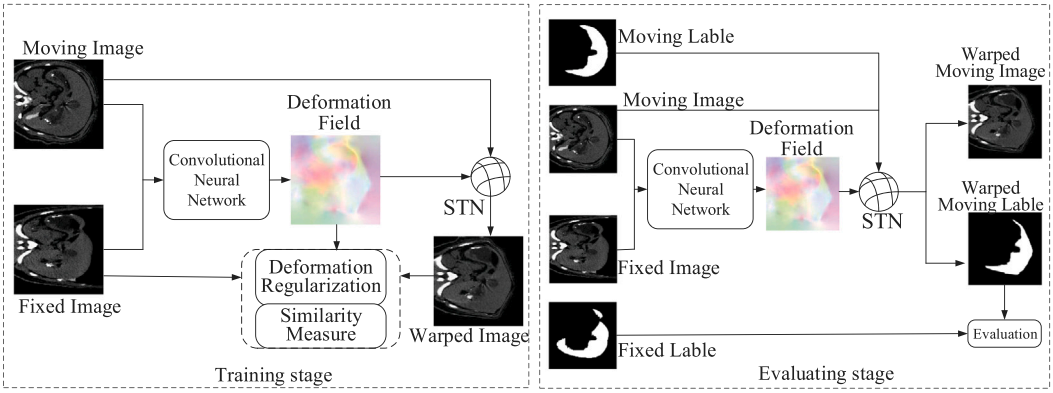

Fig. 1. The framework of our proposed unsupervised 3D medical image registration strategy.

The above observations with respect to the IVTN and VoxelMorph provide us some clues on designing an efficient convolutional neural network for our image registration tasks. The architecture of our deformable registration network (DR-Net) is shown in Fig. 2 which also appears a U-shape encoding–decoding structure. The encoding path consists of 5 Inception Blocks, 5 convolutional residual blocks (ConvRes blocks) and one pyramid input module while the decoding path contains 4 SCAM blocks, 4 Convolutional Residual Blocks and 4 up convolutional layers. In the encoding stage, features are extracted with convolutional layers while in the decoding stage the size of the feature map is recovered to the size of the original image with up convolutional layers. In our network, low level feature maps in the encoding path and their corresponding high level feature maps in the decoding path are concatenated through ResPath rather than simple skip connections as in the traditional U-Net. The size of the smallest feature map of our DR-Net is $8 \times 8 \times 8$ which is much larger than that of the IVTN. The DR-Net is used to predict a deformation field with a resolution of $128 \times 128 \times 128$ for warping the moving image to match the fixed one.

2.2.1. Convolutional residual block

The architecture of our convolutional residual block is shown in Fig. 2(d) which consists of two successive $3 \times 3 \times 3$ convolutional layers and one $1 \times 1 \times 1$ convolutional layer that enables the block fusing the feature extracted with the first $3 \times 3 \times 3$ convolutional layer and that extracted with two successive $3 \times 3 \times 3$ convolutional layers. As a basic component of our DR-Net, the convolutional residual block improves the feature reusability of the network and hence avoids using excessively deeper networks with overfitting problem.

2.2.2. Residual path

We use residual paths rather than the simple skip connections to fuse low level features from the encoder and high level features from the decoder. In terms of these paths, features from the encoder are sequentially processed with several convolutional residual blocks so that the disparity in the semantic levels between encoder and decoder features can be reduced before fusing them. As illustrated in Fig. 2, the semantic disparity between these two kinds of features becomes larger and larger when we move from bottom up and hence we use more and more convolutional residual blocks from bottom up to balance the semantic levels for features to be fused.

2.2.3. Sequential inception block (IncepBlock)

We design a light weighted sequential inception block (IncepBlock) as shown in Fig. 2(a) to enable the network focusing on registration of regions of a specific organ with different sizes and shapes. Different from the traditional GoogLeNet Inception [30] where $3 \times 3 \times 3$, $5 \times 5 \times 5$ and $7 \times 7 \times 7$ convolutional filters are used in parallel, we use a sequence of smaller and light weighted $3 \times 3 \times 3$ convolutional blocks instead to reduce the amount of training parameters. The outputs of the second and the third $3 \times 3 \times 3$ convolutional blocks effectively approximate the $5 \times 5 \times 5$ and $7 \times 7 \times 7$ convolution operations, respectively [31]. This allows us to extract spatial features from different context sizes with fewer parameters. The parallel connection mode of mean-pooling and max-pooling is beneficial to preserve more background information and image texture information, simultaneously. The introduction of the $1 \times 1 \times 1$ filter is not only for conserving dimensions but also for establishing a residual connection. Therefore, introducing inception blocks will facilitate the network to integrate the features learned from the image in different scales.

2.2.4. Pyramidal Input Module (PIM)

In the encoding stage, after each inception block, the size of the feature map will be halved, which may result in the loss of some important image information. Meanwhile, both the maximum pooling and mean pooling operations also lead to loss of important spatial information. In order to partly recover the lost spatial information and make the model better adapting to scale changes of images, we design a pyramidal input module where concatenated image pairs in 4 different scales (*i.e.*, $128 \times 128 \times 128$, $64 \times 64 \times 64$, $32 \times 32 \times 32$, $16 \times 16 \times 16$) are fed into the encoder in different levels, respectively. Here, the concatenated image pair at the first scale corresponds to the fixed and moving images to be registered while the concatenated image pair in any other scale corresponds the convolved version of the fixed and moving image (feature) pair in its previous scale with a $3 \times 3 \times 3$ kernel and with a stride of 2. Except for the first scale where the input is fed directly into an inception block, the input at any other scale is first fused with the output of the ConvRes block from the previous scale which is then fed into the Inception block to extract features in its current scale. By means of the pyramidal input module, multi-scale features of the same pair of images extracted in different ways can be combined together, which not only recovers parts of the lost information but also makes features of the moving and the fixed images more hierarchical.

2.2.5. Spatial channel attention mechanism block (SCAM block)

The attention mechanism can help the CNN learn to highlight useful voxels or channels while suppress irrelevant voxels or channels. A typical one of such attention mechanism is the squeeze and excitation (SE) network proposed in [32]. However, this network owns the following two disadvantages: (1) In the SE-block, only average pooling operations are used to compress the spatial information. But the average pooling operation can lead to blurring and loosing image details, which makes it difficult for the SE-block to well capture the rich input pattern information; (2) The SE-block only encodes inter-channel information but neglects the important positional information which is critical to capturing object structures. To resolve these two problems, we particularly design a spatial channel attention mechanism

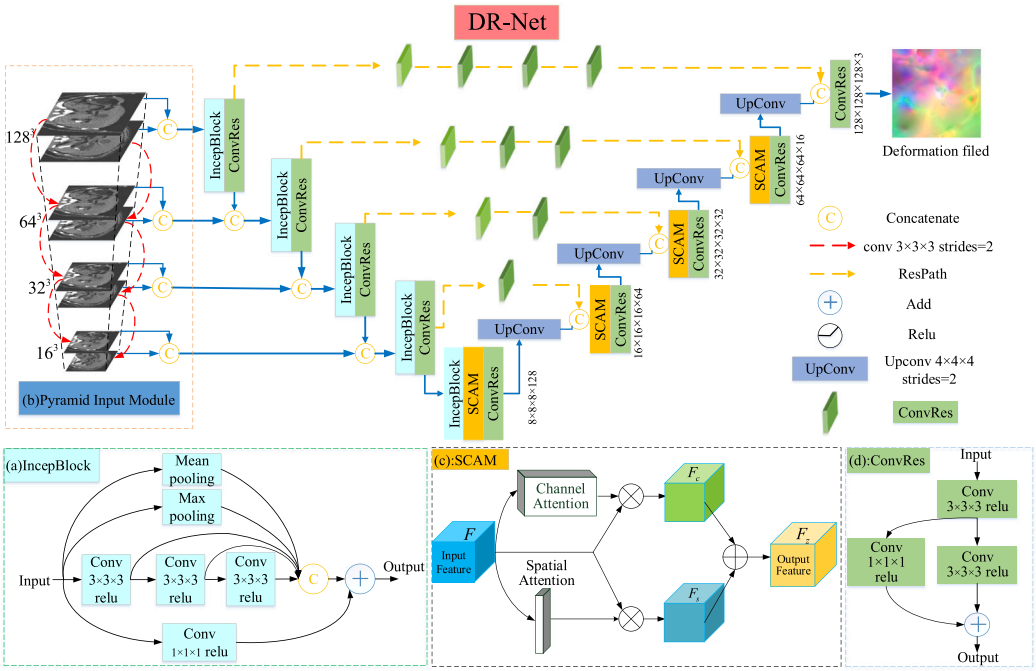

Fig. 2. The details and structure of our deformable registration Network (DR-Net). (a) plot of the IncepBlock module, (b) plot of the pyramidal input module, (c) plot of the SCAM Block module, (d) plot of the convolutional residual module.

block (SCAM block) which combines together both channel attention mechanism and spatial attention mechanism.

(A) *Channel Attention Mechanism (CAM)*

The top half of Fig. 3 gives the architecture of our Channel Attention Mechanism (CAM). To tackle the problem of information loss due to the average pooling operations in the SE-block, both the global average pooling (GAP) and the global maximum pooling (GMP) are used in parallel to aggregate the spatial information in the feature maps and hence increase the diversity of features. The GAP is used to suppress the variance increment of estimations due to the limited neighborhood size and hence better preserve the background information while the GMP is used to acquire the maximum value of the feature in a neighborhood so that the shift of the estimated mean value due to the parametric errors of the convolutional layers can be reduced. The dual-channel parallel sampling method allows the network learning to extract effective features with rich and complementary information. As illustrated in the top half of Fig. 3, the global average pooling feature $F^c_{\text{avg}} \in \mathbb{R}^{1\times1\times1\times C}$ and the global maximum pooling feature $F^c_{\text{max}} \in \mathbb{R}^{1\times1\times1\times C}$ (with C being the number of channels) are respectively fed into a shared multi-layer perceptron (MLP) with one hidden layer. In order to reduce the number of parameters, the size of the hidden layer is set as $1 \times 1 \times 1 \times C/r$. Here, r is a reduction ratio and is set as 16 in our experiments. The two new descriptors output from the MLP are then superimposed together in element-by-element manner whose result is sent to the sigmoid function to obtain weights in $(0, 1)$ which comprise the channel attention vector $M_c(F) \in \mathbb{R}^{1\times1\times1\times C}$. Finally, these weights are used to multiply the initial feature maps in their corresponding channels, which results in the following weighted feature F_c:

$$\begin{aligned} F_c &= F \times M_c(F) \\ &= F \times \{\sigma[MLP(GAP(F))] + \sigma[MLP(GMP(F))]\} \\ &= F \times \left\{\sigma\left[W_2\delta\left(W_1 F^C_{avg}\right)\right] + \sigma\left[W_2\delta\left(W_1 F^C_{max}\right)\right]\right\} \end{aligned} \tag{1}$$

where σ and δ are the sigmoid and ReLU activation functions, respectively, $W_1 \in \mathbb{R}^{C/r\times C}$ and $W_2 \in \mathbb{R}^{C\times C/r}$ are the training weights of the MLP. Here, the activation ReLU is used for the hidden layer to improve the nonlinearity of the network while the activation σ is used for the output layer to output weights in $(0, 1)$. From Eq. (1), we can see that upon using the complementary information extracted by GAP and GMP, the channel information in the input feature map is weighted to enhance important feature channels while suppress the non-important ones so that the network focuses more on those meaningful feature channels.

(B) *Spatial Attention Mechanism (SAM)*

As mentioned before, the SE block ignores issues in enabling the network learning to differentiate useful or irrelevant voxels in the same channel for image registration tasks. To tackle this problem, we additionally introduce Spatial Attention Mechanism (SAM) whose architecture is shown in the bottom half of Fig. 3. As in the CAM, global average pooling and global maximum pooling are first performed along different channels with respect to the feature map $F \in \mathbb{R}^{W\times H\times D\times C}$, which result in two different spatial descriptors $F^s_{avg} \in \mathbb{R}^{W\times H\times D\times 1}$ and $F^s_{max} \in \mathbb{R}^{W\times H\times D\times 1}$ representing the average pooling and maximum pooling features, respectively. These two features are then concatenated to construct a new feature in size of $W \times H \times D \times 2$. This new feature is subsequently sent to three successive convolutional layers whose kernel size is $3 \times 3 \times 3$, stride is 2 and padding style is the same. Upon sequentially connecting three convolutional layers with a kernel size of $3 \times 3 \times 3$, the equivalent receptive field of the last convolutional layer can be increased to $7 \times 7 \times 7$ at the expense of only increasing a fairly smaller amount of network parameters. This strategy is similar to that in our sequential inception block. It should be pointed out that the activation function of the first two layers is selected as the ReLU to improve the nonlinearity while the activation function of the last layer is chosen as the sigmoid function so that the resulted weights are in $(0, 1)$. Finally, we obtain a spatial attention vector $F_s \in \mathbb{R}^{W\times H\times D\times 1}$ which is then used to weight the initial feature map to obtain the following spatially weighted feature map F_s:

$$\begin{aligned} F_s &= F \times M_s(F) \\ &= F \times \sigma\left\{f^{3\times3\times3} f^{3\times3\times3} f^{3\times3\times3}[GAP(F), GMP(F)]\right\} \\ &= F \times \sigma\left[f^{3\times3\times3} f^{3\times3\times3} f^{3\times3\times3}\left(F^s_{avg}, F^s_{max}\right)\right] \end{aligned} \tag{2}$$

where σ is the sigmoid activation function, $f^{3\times3\times3}$ is the $3 \times 3 \times 3$ convolutional operation.

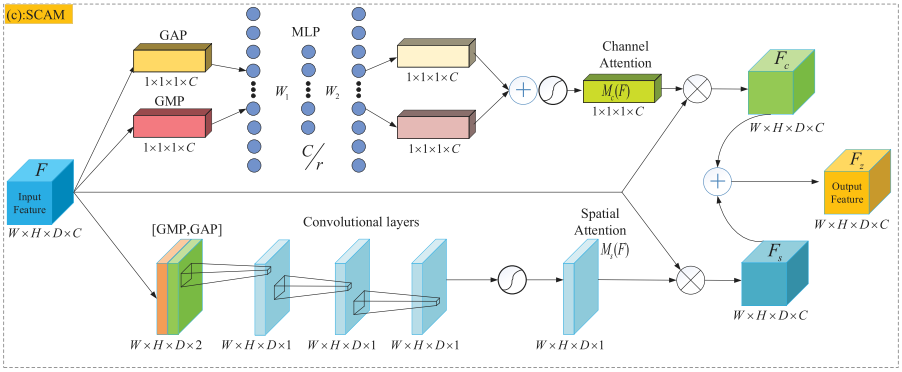

Fig. 3. The architecture of the spatial channel attention mechanism.

(*C*) *Fusion of CAM and SAM*

In order to enable the network simultaneously learning to highlight useful channels and voxels while suppressing irrelevant ones, we connect the CAM and SAM in parallel manner as illustrated in Fig. 3 to establish our SCAM block. In terms of these two branches, both the channel attention weighted feature map F_c and the spatial attention weighted feature map F_s can be obtained from the initial feature map. Superimposing them together, we can obtain the following feature map F_z which fuses features learnt from spatial domain and that learnt from channel domain:

$$F_z = F_s + F_c \tag{3}$$

As illustrated in Fig. 2, the above developed spatial-channel attention module is used after the concatenation operation to assign different weights to different channels and different regions so that the network can learn to highlight useful channels and focus on important regions while suppress irrelevant channels and regions of the concatenated features.

2.3. Multi-scale deep cooperative cascades

Similar to most other registration methods, our method also faces challenges when it is employed for registration of images with large deformation which is popular for registration tasks such as the liver registration. One way to address this problem is to stack multiple convolutional neural networks (CNNs) to constitute a cascade of deep networks for registration [26–28]. That is, the input image pair is first registered with the leading sub-network in the cascade, then the warped moving image and the fixed image are further registered with the second sub-network, and so on. This successive registration procedure enables the final prediction (probably with large displacement) to be decomposed into cascaded and progressive refinements (with small displacements). The initial sub-network is more effective for the registration of small deformed input images, while the deep cascade with more sub-networks is more helpful for the registration of image pairs with large deformation. Nevertheless, although network cascading possibly solves complex or large deformation problem, the non-cooperative and single scale training strategy prevents the cascaded network from further improvement of the registration accuracy even if more sub-networks are appended to it. Therefore, new strategies should be developed to make the model more trainable and make the registration more accurate. With this regard, we try to design a cooperative and multi-scale unsupervised training strategy for cascaded deep registration networks which can perform both cooperative training between different sub-networks in the same scale and multi-scale cooperative training within the sub-network itself and learn to perform progressive alignments cooperatively. Fig. 4 illustrates an example of our designed cascade of two cooperative DR-Nets that mainly includes two parts: multi-scale cooperation within each sub-network and the cooperation between cascaded sub-networks. From Fig. 4, we can see that our strategy is quite different from the existing non-cooperative cascades where the successive sub-networks are connected with successively warped images, we focus not only on exploiting the hierarchical information contained in different layers of the sub-networks themselves but also on the interrelationship between sub-networks and try to achieve better registration accuracy with a relatively shallower cascade.

(*A*) *Multiscale cooperation within the network*

Most registration methods calculate deformation field only from the feature map of the last layer in the network. However, this strategy does not make the best use of the hierarchical information contained in feature maps of different layers of the network for predicting the deformation fields. To tackle this problem, we propose a cooperative strategy within each of the DR-Net in the cascades where the deformation field predicted in any shallow layer is incorporated with that predicted in the subsequent deeper layer to generate an accumulated deformation field. This mechanism is helpful to mine and utilize the hierarchical information contained in convolutional layer of different scales within the sub-network so that small parts of different scales in largely deformed images can be accurately registered.

According to our multiscale cooperative strategy, the accumulated deformation field ϕ_n^{a,l_k} in the kth scale of the nth sub-network can be recursively obtained as:

$$\begin{aligned} &\phi_n^{a,l_k} = \mathbf{A}\left(\phi_n^{l_k}, \varphi_n^{a,l_{k-1}}\right) \\ &\phi_n^{a,l_2} = \phi_n^{l_2} \\ &n = 1, 2, \ldots; k = 3, 4 \end{aligned} \tag{4}$$

where $\mathbf{A}$ denotes an accumulated operator, $\phi_n^{l_k}$ is the deformation field predicted in scale l_k of the nth cascaded sub-network, $\varphi_n^{a,l_{k-1}}$ denotes the up-convolved version of the accumulated deformation field $\phi_n^{a,l_{k-1}}$ from scale l_{k-1} to scale l_k with a kernel size of $4 \times 4 \times 4$ and a stride of 2.

(*B*) *Multi-scale Cooperation between cascaded sub-networks*

Fig. 4 illustrates our cooperative cascade of the DR-Nets for successive registration of largely deformed image pairs where the relationship between two successive sub-networks is established not only in terms of the successively warped moving images but also in terms of the connections between corresponding layers of two successive sub-networks, which is quite different from traditional cascades whose sub-networks are interrelated only with successively warped moving images. In this way, we can integrate the deformation fields predicted in different layers of the proceeded sub-network to the corresponding ones predicted in different layers of the succeeded sub-network whose results are then aggregated together to generate a cooperative deformation field to produce a warped moving image input to the next sub-network.

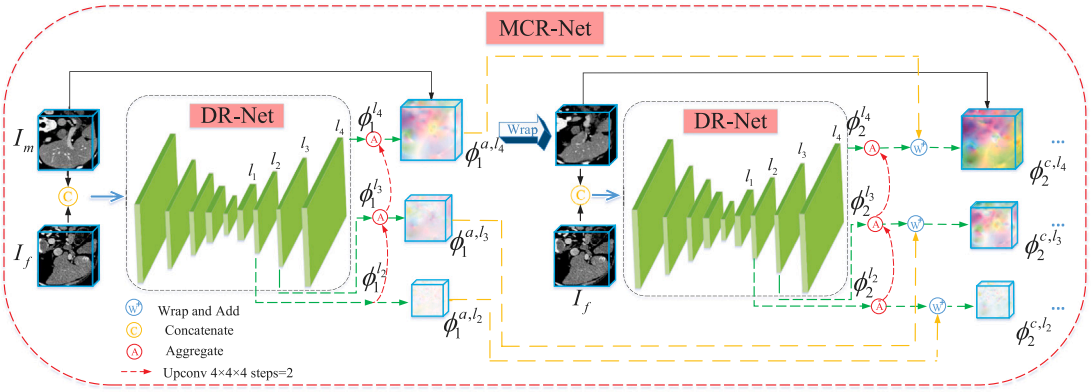

Fig. 4. Schematic illustration of the multi-scale cascading registration network.

According to our inter-sub-network cooperation strategy in our cascade, the final cooperative deformation field for different cascaded sub-networks and in different scales can be recursively expressed as:

$$\begin{aligned}\phi_1^{c,l_k} &= \phi_1^{a,l_k}\\ \phi_n^{c,l_k} &= \mathbf{w}^{+}\left[\phi_n^{a,l_k}, \phi_{n-1}^{c,l_k}\right]\\ &= \mathbf{w}\left[\phi_{n-1}^{c,l_k}, \phi_n^{a,l_k}\right] + \phi_n^{a,l_k}\\ n &= 2,3,\ldots,k = 2,3,4\end{aligned} \tag{5}$$

where ϕ_n^{c,l_k} is the cooperative deformation field of the nth sub-network in scale l_k, $\mathbf{w}^{+}$ is a composite operator including the warping and adding operations, $\mathbf{w}(\cdot,\cdot)$ is the warping operator which performs the warping operation with respect to the first argument according to the second one. With the cooperative deformation field, the registered image $I_{m,r}$ of I_m can be computed as:

$$I_{m,r} = \mathbf{w}\left(I_m, \phi_{n_l}^{c,l_4}\right) \tag{6}$$

where $\phi_{n_l}^{c,l_4}$ is the cooperative deformation field in scale l_4 of the last cascaded sub-network. From Eq. (6), we can see that the final deformation field used to warp the moving image is the output of layer l_4 in the last sub-network. After the vertical fusion within each sub-network and the horizontal fusion between different sub-networks, multi-scale displacement information has been fused in the final cooperative deformation field $\phi_{n_l}^{c,l_4}$ of the last sub-network, which will be extremely helpful for the registration of largely deformed image pairs.

2.4. Multi-scale loss function

In most existing cascading methods, the training process is only supervised by the training loss calculated in the final layer of the network, which often results in suboptimal estimation of parameters in the frontal convolution layers because parameters of the frontal convolution layers are not updated as much as those of the latter part. This not only causes slow convergence but also causes the over-fitting problem. To resolve these problems, we propose a multiscale training strategy where not only the training loss calculated in the final decoder layer of the last sub-network but also those calculated in other decoder layers of the last sub-network are used to supervise the training process. A schematic illustration of this strategy is shown in Fig. 5 where the similarity between the warped moving images and the fixed images are computed in l_2, l_3 and l_4 layers of the last sub-network. There are some advantages with this training strategy: (1) With the multi-scale loss, the training of the first (frontal) half layers of the decoder can be supervised directly and hence the training process can be speeded up; (2) As mentioned before, the deformation field ϕ^{c,l_2}, ϕ^{c,l_3}, ϕ^{c,l_4} output from l_2, l_3 and l_4 layers contains multi-scale displacement details and hence the introduction of multi-scale loss function computed in terms of ϕ^{c,l_2}, ϕ^{c,l_3}, ϕ^{c,l_4} will enable the network learning to better register details in medical images hierarchically. From Fig. 5, we can see that the spatial resolution of the input image pairs is $128 \times 128 \times 128$ while the spatial resolution of the output deformation field in l_2, l_3 and l_4 layers are $32 \times 32 \times 32$, $64 \times 64 \times 64$ and $128 \times 128 \times 128$, respectively. Due the scale difference between the predicted deformation field (in l_2, l_3 layers) and the input moving images, we need to up sample the deformation field to make their spatial resolution consistent with each other (i.e., up sample the deformation field output in l_2, l_3 layers to $128 \times 128 \times 128$). Here, the bilinear interpolation operations are implemented to up sample ϕ^{c,l_2} and ϕ^{c,l_3} into Φ^{l_2} and Φ^{l_3} with a resolution of $128 \times 128 \times 128$, respectively. Besides the similarity loss between the warped moving images and the fixed ones, regularization terms are also introduced in each scale to prevent the deformation fields from being unrealistic or overfitting. Another loss term used here is the edge loss introduced to enable the network learning to better match the structures in medical images. Therefore, the loss supervising the training process of our model consists of three parts: (1) $L_{sim}\left[I_f, \mathbf{w}(I_m, \Phi)\right]$-the similarity measure between the fixed image I_f and the moving image I_m warped according to the deformation field Φ currently estimated via the network; (2) $L_R(\Phi)$-the regularization term to prevent the unreasonable deformation field Φ predicted in the registration neural network; (3) $L_{edge}\left[I_f, \mathbf{w}(I_m, \Phi)\right]$-the edge loss measuring the extent of mismatch between the edge images of the warped moving image $\mathbf{w}(I_m, \Phi)$ and of the fixed image I_f.

2.4.1. The similarity loss

Upon introducing the similarity measure in the loss function, the training process forces the network updating its weights so that its predicted flow field can deform the moving image to a version most similar to the fixed image in terms of the similarity measure. Candidates for such a measure in our mono-modal medical image registration tasks include the mean square error (MSE), the correlation coefficient (CC), the normalized correlation coefficient (NCC). Here, we use CC as our similarity metric which has been widely used in registration networks (such as IVTN and VoxelMorph networks). The correlation coefficient between the fixed image I_f and the warped moving image $I_m^k = \mathbf{w}(I_m, \Phi^k)$ in the kth ($k = l_2, l_3, l_4$) scale is defined as:

$$CC[I_f, I_m^k] = \frac{Cov[I_f, I_m^k]}{\sqrt{Cov[I_f, I_f]Cov[I_m^k, I_m^k]}} \tag{7}$$

where Φ^k denotes the up-sampled (to input scale) deformation field of $\phi^{c,k}$ in the kth scale, $\mathbf{w}$ is a warping operator and

$$Cov[I_f, I_m^k] = \frac{1}{|\Omega|}\sum_{i\in\Omega} I_f(i) I_m^k(i) - \frac{1}{|\Omega|^2}\sum_{i\in\Omega} I_f(i) \sum_{j\in\Omega} I_m^k(j) \tag{8}$$

is the covariance between I_f and I_m^k with Ω denoting the spatial domain occupied by the 3D images, $|\Omega|$ being the number of voxels

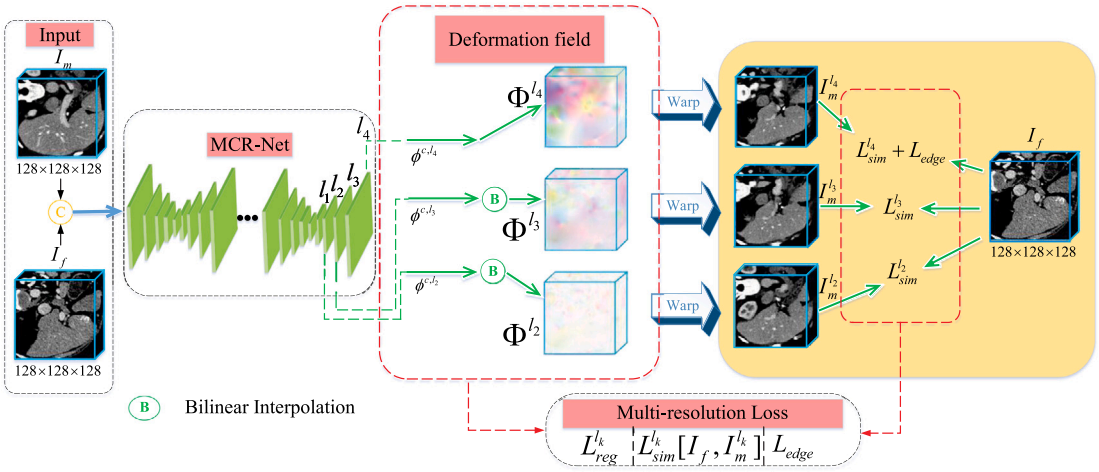

Fig. 5. Schematic illustration of multi-resolution loss function construction.

included in Ω, both i and j being the voxels in Ω. Since the value of the correlation coefficient lies in $[-1, 1]$, to obtain a non-negative value, the similarity measure in the loss function is defined as follows:

$$L_{sim}^{k}[I_f, I_m^k] = 1 - CC[I_f, I_m^k], k = l_2, l_3, l_4 \tag{9}$$

2.4.2. Regularization loss

For a dense flow (deformation) field Φ^k which is up-sampled from ϕ^k in the kth scale, we regularize it with the following loss that discourages discontinuity:

$$L_R^k = \frac{1}{3|\Omega|} \sum_x \sum_{i=1}^{3} \left(\Phi^k(\mathbf{x} + \mathbf{e}_i) - \Phi^k(\mathbf{x})\right)^2, k = l_2, l_3, l_4 \tag{10}$$

where $\mathbf{e}_{1,2,3}$ is the orthogonal basis of $\mathbb{R}^3$, L_R^k is our regularization loss term in the kth scale. Regularization loss will ensure that the image distorted by the predicted deformation field will maintain smooth and realistic, which is of great importance to the registration performance. This loss will be applied to all registration fields used to deform the moving image.

2.4.3. Edge loss

The similarity loss in Eq. (9) focuses on the voxel-level similarity degree between two images while ignore the image details, such as the texture or shape of an organ. The edge information has the capability to describe the contour of the image and hence it will be helpful for the registration of image details if the edge information is used to guide the model training. Therefore, we incorporate the edge similarity loss in the loss function to enable the network learning to better align image structures. Here, the 3D Sobel operator designed by Yu [33] is utilized to extract the edge information of the image pair. This operator includes three 3D kernels (F_x, F_y, F_z) corresponding to x, y and z directions, respectively. The structure of these 3D operators in x, y and z directions are shown in Fig. 6. With these three Sobel kernels, the edge information for any image I can be obtained as in Eq. (11).

$$S(I) = \sqrt{(F_x * I)^2 + (F_y * I)^2 + (F_z * I)^2} \tag{11}$$

where $*$ is the convolution operator. Based on the edge information in Eq. (11), the edge loss is defined as follows:

$$L_{edge}^k = \frac{1}{|\Omega|} \left\| S(I_f) - S(I_m^k) \right\|_2, k = l_2, l_3, l_4 \tag{12}$$

2.4.4. Multiscale integrated loss (MIL)

Combining the regularization loss, the edge loss and the similarity loss function together, we can obtain the final multiscale integrated loss (MIL) function for our multiscale training task as follows:

$$L_{integrate} = \lambda_1 L_{sim} + \lambda_2 L_{edge}^{l_4} + L_R \tag{13}$$

where λ_1 and λ_2 are two combinatorial coefficients and

$$\begin{aligned} L_{sim} &= L_{sim}^{l_4}[I_f, I_m^{l_4}] + \alpha L_{sim}^{l_3}[I_f, I_m^{l_3}] \\ &\quad + \beta L_{sim}^{l_2}[I_f, I_m^{l_2}] \end{aligned} \tag{14}$$

and

$$L_R = L_R^{l_4}(\Phi^{l_4}) + \alpha L_R^{l_3}(\Phi^{l_3}) + \beta L_R^{l_2}(\Phi^{l_2}) \tag{15}$$

Here, α, β are two non-negative coefficients. As before, $I_m^k = \mathbf{w}(I_m, \Phi^k)$ denotes the warped moving image of I_m with the up-sampled deformation filed Φ^k. Due to blur degradation of the image with low spatial resolution, the edges extracted in the blurred image cannot properly describe the structure of the organs in medical images and hence the edge loss in our work is only computed in the upmost layer l_4 with highest resolution of 128 × 128 × 128. In terms of the multi-scale integrated loss function and multi-layer registration scheme from high, middle and low resolutions, the multi-scale deformation fields output from different layers of the last sub-network but integrate the deformation information of other sub-networks in the cascade are used to compute the integrated loss governing the training process.

During the training process, we use the multi-scale integrated loss of the last sub-network in the cascade to supervise the multi-scale training of the whole cascade cooperatively. Theoretically, we can achieve better results in terms of deeper cascades. But here we lay more emphasis on the cooperation in between sub-networks and within the sub-networks themselves to achieve comparable or even better results with shallower cascades, which not only eases the hardware requirements but also reduces the amount parameters and speeds up the registration process.

3. Experiments

3.1. Datasets

Our experiments are conducted on four datasets of 3D CT images of the liver from different medical centers.

- MSD [34] is a dataset that contains 10 types of medical images for different semantic segmentation tasks. The dataset includes 933 scans of 3D liver parts without segmentation label.

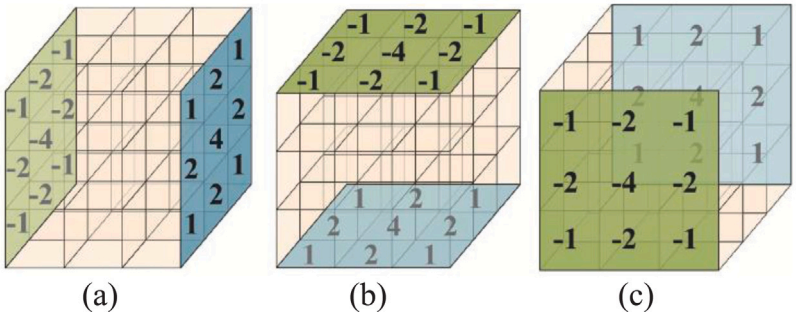

Fig. 6. Schematic illustration of the 3D Sobel kernel with size of 3 × 3 × 3.

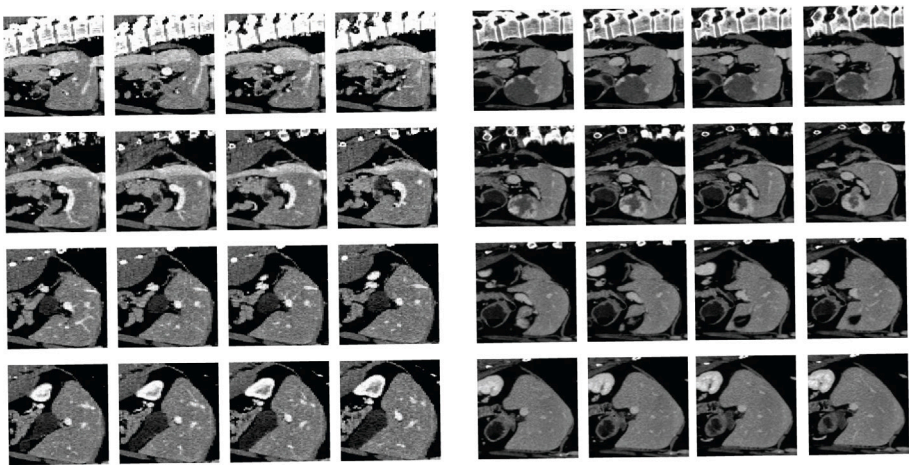

Fig. 7. Sample images from the BFH dataset (left) and the Sliver dataset (right).

- BFH [27] consists of 92 liver volumes provided by Beijing Friendship Hospital, China.
- LiTS [28] is a dataset provided by different clinical sites worldwide which contains 131 3D CT scans with the ground truth of liver segmentation.
- Sliver [35] contains 20 3D CT scans with the ground truth of liver segmentation. In these scans, 6 anatomical key points annotated by 3 doctors are selected as landmarks, and their average value for each landmark is taken as ground truth.

In our experiment, totally 1025 scans of 3D images from the MSD and BFH datasets are employed for training the unsupervised registration network while the remaining 151 scans (131+20) of 3D images from the LiTS, Sliver datasets are used for evaluation. The above four datasets are all publicly available datasets. Sample images from BFH and Sliver datasets are shown in Fig. 7.

3.2. Experimental settings

- Implementation
  Our model is trained and implemented on the deep-learning framework of Keras 2.3.1 and TensorFlow 1.13.1. We train our model on 1 card of 32G NVIDIA TESLA V100 32G GPU and CUDA 10.1, Cudnn 7.4.1. The CPU is Intel(R) Xeon(R) Gold 6240R CPU @ 2.40 GHz.
- Baseline Methods
  In order to evaluate our proposed method comprehensively, we conduct some comparative experiments. The results of our proposed method are compared with two popular deep learning methods for medical image registration including VoxelMorph [13] and IVTN [28].

3.3. Evaluating metrics

The following three metrics are employed to evaluate the performance of different registration methods:

- Dice Score
  Dice Score is a metric to measure the extent of overlapping of two samples. Here, we use the Dice Score to measure the overlap of the liver regions in the fixed and registered moving images. Suppose that S_f is the liver region in the fixed image and S_m is the liver region in the registered moving image, the Dice Score between S_f and S_m is calculated as:

$$Dice(S_f, S_m) = 2\frac{|S_f \cap S_m|}{|S_f| + |S_m|} \tag{16}$$

  In our evaluation, the Dice Score is computed based on the ground truth of liver segmentation. Obviously, the better the moving image is registered to the fixed image, the larger the Dice Score will be.
- Similarity
  The similarity is measured by the correlation coefficient between the registered moving image and the fixed image as defined in Eq. (7). A value of 1 implies totally identical, a value of 0 implies completely unrelated, while a value of −1 implies that one image is a negative version of the other.
- The amount of parameters
  The amount of parameters included in the model is a key factor affecting the hardware requirement for both training and testing. Meanwhile, light models are extremely helpful for avoiding the over-fitting problem especially for medical image processing tasks where only limited amount of annotated samples are available.

Table 1
Performance (± std) of the DR-Net with different configurations on the Sliver dataset.

Deep models	IncepBlock	PIM	SCAM	*Dice*	*Sim*
1	×	×	×	0.898 ± 0.029	0.765 ± 0.038
2	√	×	×	0.903 ± 0.027	0.775 ± 0.038
3	√	√	×	0.909 ± 0.024	0.782 ± 0.033
4	√	√	√	$\mathbf{0.915 \pm 0.024}$	$\mathbf{0.796 \pm 0.030}$

Table 2
Performance (± std) of the DR-Net with different configurations on the LiTS dataset.

Deep models	IncepBlock	PIM	SCAM	*Dice*	*Sim*
1	×	×	×	0.852 ± 0.053	0.754 ± 0.050
2	√	×	×	0.858 ± 0.052	0.763 ± 0.048
3	√	√	×	0.863 ± 0.052	0.772 ± 0.048
4	√	√	√	$\mathbf{0.871 \pm 0.052}$	$\mathbf{0.782 \pm 0.042}$

3.4. Training and optimization

In our training stage, the learning rate is adjusted adaptively according to the following equation:

$$lr = lr\_base \times \left(\frac{1}{2}\right)^{\left\lfloor \frac{ecoph}{10000} \right\rfloor} \tag{17}$$

where lr is the learning rate of training, lr_base is the initial training rate (e^{-4} in our experiment), $epoch$ is the count of the current training rounds and totally 30,000 rounds of training are implemented, $\lfloor x \rfloor$ is a floor operator rounding down x to a maximum integer less than it. From Eq. (17), we can see that every 10,000 rounds of training, the learning rate will be halved. In our experiments, the batch size is set as 1.

4. Results

In our experiment, before deformable registration, the rough registration is first performed amongst the image pair to be registered with the affine registration network proposed in [28] and trained on the MSD and BFH datasets.

4.1. Ablation analysis of different modules and edge loss in DR-Net

In order to investigate the impact of different modules (PIM, IncepBlock and SCAM) of the DR-Net on its registration performance, ablation experiments are performed. In all cases, deep models with different configurations are first trained on the MSD and BFH datasets and then tested on the LiTS, Sliver datasets whose results are shown in Tables 1 and 2, respectively. It should be pointed out that in these ablation experiments we use the single scale training loss $L_{integrate} = L_{sim}^{l_4} + L_R^{l_4}$ computed according to Eqs. (9), and (10) with $k = l_4$ of the DR-Net as illustrated in Fig. 2. From these two tables, we can see that when more and more modules are incorporated in the DR-Net, the performance becomes better and better on both test datasets. Comparing results from model 2 with those from model 1 in Table 1, we can see that, upon incorporating the IncepBlock in the DR-Net, we can achieve an improvement of 0.5% in the Dice Score and 1% in the similarity measure on the Sliver dataset. It can also be observed from Table 2 that such improvements amount to 0.6% and 0.9% on the LiTS dataset. Experimental results demonstrate that the PIM can further improve the registration accuracy amounting to 0.6% and 0.7% on the Sliver dataset while 0.5% and 0.9% on the LiTS dataset. Comparing model 3 with model 4, we can observe that the attention mechanisms (SCAM) is helpful for improving the registration accuracy in terms of both the Dice Score and the Similarity measure after they are incorporated in our DR-Net, which amounts to 0.6% and 1.4% on the Sliver dataset while 0.8% and 1.0% on the LiTS dataset. As compared with model 1 without any module incorporated, we achieve 1.7% and 3.1% improvements on the Sliver dataset while 1.9% and 2.8% improvements on LiTS dataset in terms of the Dice Score and Similarity measure when the PIM, the SCAM and the IncepBlock are all introduced in the DR-Net. Pair-wise statistical t-tests performed in between models 1 and 2, models 2 and 3, models 3 and 4 demonstrate that the improvements achieved by successively introducing the IncepBlock, the PIM and the SCAM block are of statistical significance ($p < 0.05$).

Table 3
Performance (± std) of the DR-Net with different loss on the Sliver dataset.

Loss	λ_1	λ_2	*Dice*	*Sim*
1	1	0	0.915 ± 0.024	0.796 ± 0.030
2	1	1	$\mathbf{0.920 \pm 0.021}$	$\mathbf{0.806 \pm 0.028}$
3	1	0.5	0.917 ± 0.024	0.801 ± 0.039
4	0.5	1	0.914 ± 0.024	0.780 ± 0.039

In order to investigate the impact of the edge loss on the registration accuracy, the performance of different models trained with a loss similar to that in Eq. (13) but with each term computed at a single scale l_4 of the DR-Net according to Eqs. (9), (10) and (12) under four different settings of λ_1, λ_2 is evaluated whose results are shown in Table 3. It can be observed that the edge loss can help to improve the registration accuracy under a proper setting of λ_1, λ_2 and the model performs the best when we set $\lambda_1 = \lambda_2 = 1$. Pair-wise statistical t-tests performed in between the first two cases in Table 3 show that the improvement in the registration accuracy due to the introduction of the edge loss is statistically significant. According to the observation here, we will set $\lambda_1 = \lambda_2 = 1$ in Eq. (13) in all the experiments hereafter.

4.2. Comparison of our DR-Net with other state-of-the-art deep models

As mentioned before, we train our proposed DR-Net with samples from datasets of BFH and MSD while evaluate it in other two datasets (Sliver, LiTS). Table 4 summarizes the performance of two state-of-the-art methods and that of our DR-Net under two different training strategies (with and without edge loss) in terms of the Dice Scores and the Similarity. It should be noted that evaluated results for IVTN and VoxelMorph models are obtained by implementing the open source codes provided by the authors under the same settings of hyper parameters for credible comparisons. Deformable registration results in Table 4 are obtained from different deep models with the affine registration results as inputs. Statistical t-tests show that our DR-Net significantly ($p < 0.05$) outperforms IVTN and VoxelMorph in two test datasets. For example, we achieve 2% and 1.9% improvements in Dice while about 9.3% and 8.9% improvements in Similarity as compared with those of the IVTN on the Sliver and LiTS datasets. As compared with the VoxlMorph on the other hand, these improvements amount to 1.9% and 2.0% in Dice while about 2.0% and 2.1% in Similarity on the Sliver and LiTS datasets. The performance of our DR-Net when the edge loss is incorporated to supervise the training process (DR-Net(L)) is shown in the last row of Table 4. Statistical t-tests also show that further significant ($p < 0.05$) improvements can be achieved in all two test datasets when the edge loss is introduced for training. For example, we achieve additional 0.5% and 0.5% improvements in Dice while additional 1% and 0.9% improvements in Similarity as compared with the DR-Net trained without the edge loss incorporated on the Sliver dataset. All the above experimental results demonstrate that our proposed DR-Net outperforms the existing deep models significantly whenever the edge loss are incorporated or not for model training. Besides the Dice Score and the Similarity measure, the performance is also evaluated in terms of the amount of parameters involved in different models whose results are shown in the last column of Table 4. From this table, we can see that our DR-Net not only achieves better registration accuracy than that of the IVTN but also contains only

Table 4
Performance ($\pm std$) of different models in terms of Dice and Similarity.

Methods	Sliver		LiTS		Para. (Mb)
	Dice	Similarity	Dice	Similarity	
Affine	0.787 ± 0.040	0.405 ± 0.065	0.751 ± 0.059	0.412 ± 0.068	162
VoxelMorph	0.896 ± 0.027	0.776 ± 0.037	0.851 ± 0.053	0.761 ± 0.047	165
IVTN	0.895 ± 0.029	0.703 ± 0.043	0.852 ± 0.053	0.693 ± 0.051	485
DR-Net	0.915 ± 0.024	0.796 ± 0.030	0.871 ± 0.052	0.782 ± 0.042	417
DR-Net(L)	**0.920 ± 0.021**	**0.806 ± 0.028**	**0.876 ± 0.050**	**0.791 ± 0.039**	417

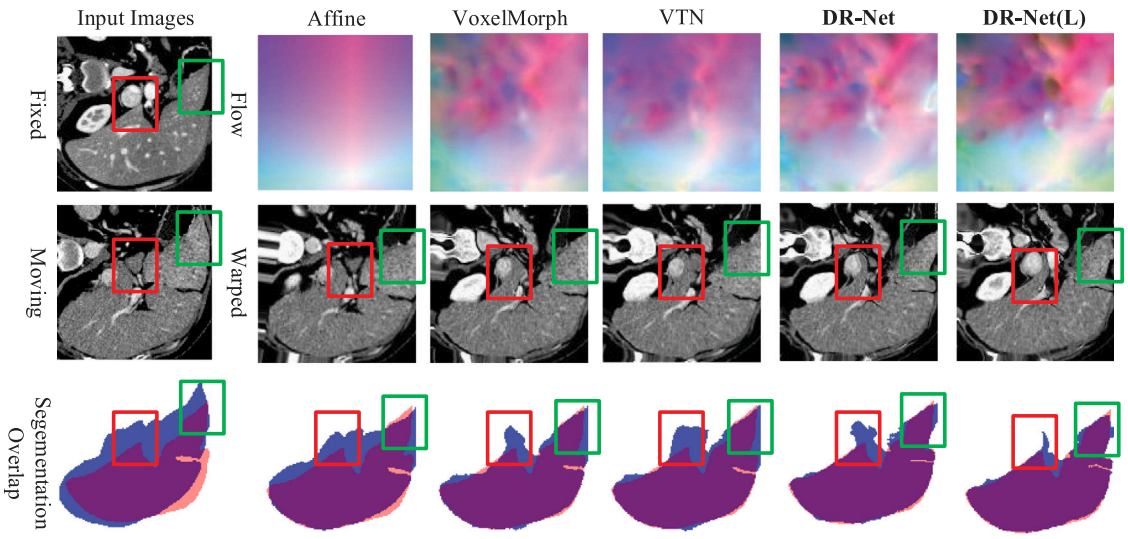

Fig. 8. Registration results from different methods for a pair of images from the Sliver dataset.

86% amount of parameters of the IVTN, which implies a significant reduction in hardware requirement. On the other hand, although our DR-Net contains more parameters than the VoxelMorph, the former performs much better in registration accuracy than the latter. We can also see from this table that the introduction of the edge loss in our DR-Net can achieve significant improvement in registration accuracy without introducing additional parameters.

For a more intuitive evaluation of the registration results, Figs. 8 and 9 give two examples of registration results from different methods for image pairs respectively from the Sliver and LiTS datasets where corresponding fixed and moving slices, deformation fields and overlapped region of livers in the fixed and warped moving images are presented. The first row gives the fixed image and the deformation fields predicted with different deep models. Here, the deformation field is demonstrated by mapping the normalized displacement in x, y and z directions to three color channels. It can be seen from these deformation fields that the deformation field predicted by our proposed DR-Net(L) contains the most abundant colors and provides more deformation details in the whole image. The second row gives the moving image and its warped versions according to the above deformation fields resulted from different deep models. Red and blue boxes highlight the difference in the registration results from different methods, which shows that all of the five models can force the moving image more similar to the fixed image but our DR-Net(L) can deform the moving image best matching the fixed image especially its fine details. Here, we focus on the local registration of liver regions rather than the whole CT volumes. Therefore, the registration accuracy for liver region is of particular importance. For visual assessment of this issue, we consider the extent of overlapping between the warped liver region in the moving image and that in the fixed image. To do so, we use the deformation fields from different methods to warp labeled liver region of the moving image and overlap them to the labeled liver region of fixed images whose results are presented in the last rows of Figs. 8 and 9 where blue areas represent the labeled liver region in the fixed image while red areas represent the labeled liver region or the warped liver regions in the moving image. Coincident areas are shown in purple while significant difference between the results from different methods are highlighted with green boxes. It can be intuitively seen that the DR-Net(L) method has the highest degree of label overlapping, which implies a best performance of the DR-Net(L) method among others.

4.3. Comparison among different cascading strategies

In order to evaluate the performance of different cascading strategies, extensive registration experiments are conducted in between image pairs from different datasets with the cascaded networks established based on the VoxelMorph, IVTN and our DR-Net. Table 5 gives quantitative evaluation of different cascades in terms of the Dice Scores and the Similarities where VM denotes the VoxelMorph, CR represents the cascading strategy as in [27], MCR represents multi-scale cascading strategy proposed in this paper, n denotes the number of sub-networks in the cascade. As expected, both cascading strategies improve the registration accuracy in terms of both the Dice Scores and the Similarities for all cascades based on the VoxelMorph, IVTN or our DR-Net but our multi-scale cascading strategy MCR outperforms the CR strategy in all cases. For example, if cascades are based on the VoxelMorph, our multi-scale cascading strategy MCR can achieve 0.7% and 0.5% more improvements in Dice Score while 0.8% and 1.2% more improvements in Similarity than the CR cascading strategy for image pairs in the Sliver and LiTS datasets, respectively. Similar improvements can be observed if the cascaded model is constructed based on the IVTN, which shows that the multi-scale cooperation cascading strategy is very helpful for improving the registration accuracy. From Table 5 we can also see that if the multi-scale cooperative cascading strategy MCR is used, cascade based on our DR-Net performs better than cascades based on the VM and the IVTN. For example, the Dice Score of the DR-MCR2 is 0.8% and 0.7% higher than those of the VM-MCR2 and IVTN-MCR2 on the Sliver dataset. Although the Similarity of our DR-MCR2 is a little bit

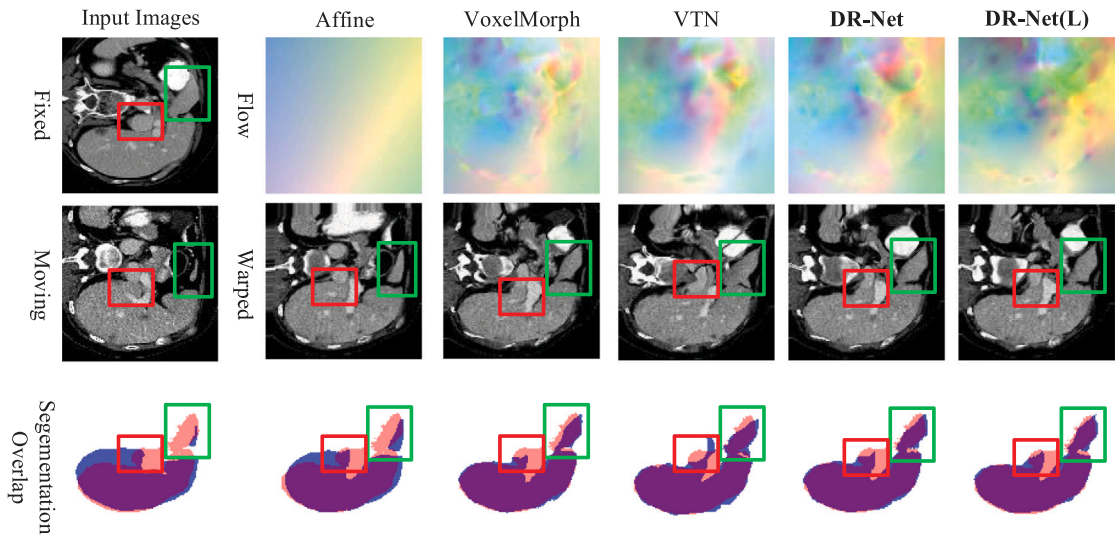

Fig. 9. Registration results of a sample pair in the LiTS dataset from different methods.

Table 5
Performance ($\pm std$) of different cascading strategies in terms of Dice and Similarity.

Methods	Sliver		LiTS		Para. (Mb)
	Dice	Similarity	Dice	Similarity	
VM	0.896 ± 0.027	0.776 ± 0.037	0.851 ± 0.053	0.761 ± 0.047	165
VM-CR2	0.917 ± 0.025	0.852 ± 0.024	0.873 ± 0.050	0.839 ± 0.035	168
VM-MCR2	0.924 ± 0.022	0.860 ± 0.024	0.878 ± 0.051	0.851 ± 0.032	169
IVTN	0.895 ± 0.029	0.703 ± 0.043	0.852 ± 0.053	0.693 ± 0.051	485
IVTN-CR2	0.915 ± 0.025	0.797 ± 0.032	0.873 ± 0.049	0.785 ± 0.041	808
IVTN-MCR2	0.925 ± 0.022	0.806 ± 0.032	0.880 ± 0.048	0.798 ± 0.039	809
DR	0.915 ± 0.024	0.796 ± 0.030	0.871 ± 0.052	0.782 ± 0.042	417
DR-CR2	0.927 ± 0.025	0.825 ± 0.027	0.885 ± 0.050	0.827 ± 0.034	668
DR-MCR2	**0.932 ± 0.022**	**0.833 ± 0.027**	**0.889 ± 0.049**	**0.828 ± 0.035**	674

lower than that of the VM-MCR2 on the Sliver dataset, it is 2.7% higher than that of the IVTN-MCR2. We can have similar observation on the LiTS dataset where the Dice Score of our DR-MCR2 is 1.1% and 0.9% higher than those of the VM-MCR2 and IVTN-MCR2. The Similarity of the DR-MCR2 is a little bit lower than that of the VM-MCR2 but is 3.0% higher than that of the IVTN-MCR2. Comparing the amount of parameters of the CR cascaded models and MCR cascaded models, we can conclude that the MCR models achieve significantly larger improvements in registration accuracy only at an expense of increasing smaller amount of model parameters.

To intuitively investigate the performance of different cascades especially in registering image pairs with large deformation, registration results from different methods for two image pairs respectively from the Sliver and LiTS datasets are presented in Figs. 10 and 11 where the first rows give the fixed images and the deformation fields predicted with different cascades while the second rows give the moving images and their warped versions according to the above deformation fields resulted from different cascaded models. The third rows give the extent of overlapping between the liver (or warped liver) regions in the moving images and those in the fixed images. As can be observed from the third rows of Figs. 10(a) and 11(a), the initial overlapping rates of liver regions in the fixed and moving images before registration are very low, which implies that larger deformation is needed to register livers in the fixed and moving images. We can also see from the first rows in these two figures that the deformation fields predicted by the MCR models contain richer colors with clearer contours and hence with more displacement details than those predicted by the CR models especially in local regions such as those highlighted with black squares, which implies a fusion of multi-scale local displacement information in the deformation fields predicted by the MCR cascades. This issue is critical for accurate registration of local areas in image pairs with large deformation. Such an inference is verified by the difference in the corresponding areas highlighted with red boxes in the third rows of overlapped liver regions. Upon inspecting areas highlighted with black boxes in the deformation fields in the first rows and their corresponding areas highlighted with red boxes in the third rows, we can see that areas with more local displacement details in the deformation fields can result in areas with higher rate of local overlapping in the corresponding areas in registered liver regions, demonstrating more accurate local registration in these areas. Comparing the registration results from different cascading strategies in the second rows, we can find that the warped moving images resulted from the deformation field predicted with the MCR models look more similar to the corresponding fixed images than those with the CR models. Such superiority of our MCR cascading strategy can also be observed from the third rows where liver regions in warped moving images from the MCR models better overlap the liver regions in the fixed image than those from the CR models do. Furthermore, liver regions in the warped moving images from our DR-MCR2 almost totally overlap the liver regions in fixed images from both datasets, which is much better than any results from other models. All the above observations demonstrate that our multi-scale cascading strategy owns particular superiority especially for registering image pairs with large deformation. It can fuse displacement information in different scales for accurate registration of small local areas in images.

4.4. Performance impact of multi-scale loss

In order to investigate the impact of the multi-scale loss on the registration accuracy, extensive experiments are conducted with multi-scale cooperative cascaded models based on our DR-Net and trained

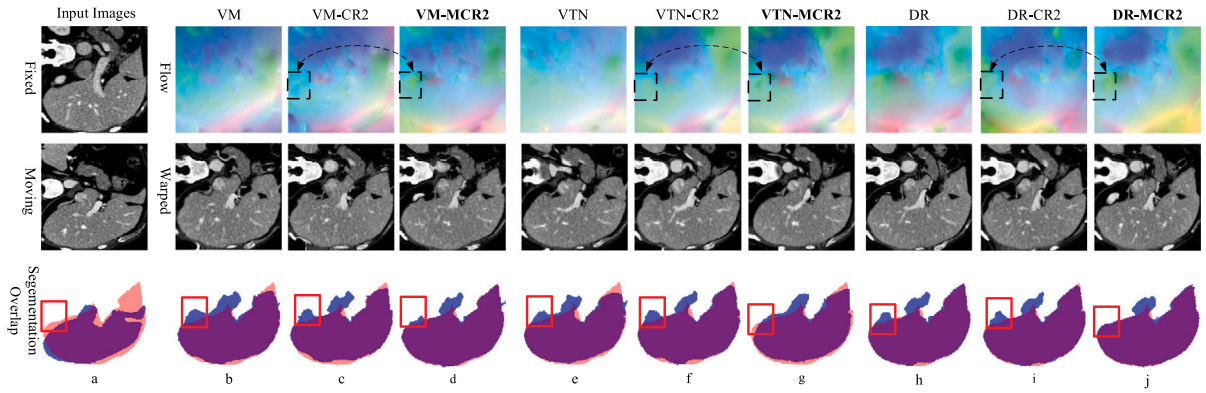

Fig. 10. Registration results from cascades with different strategies for a pair of images from the Sliver dataset.

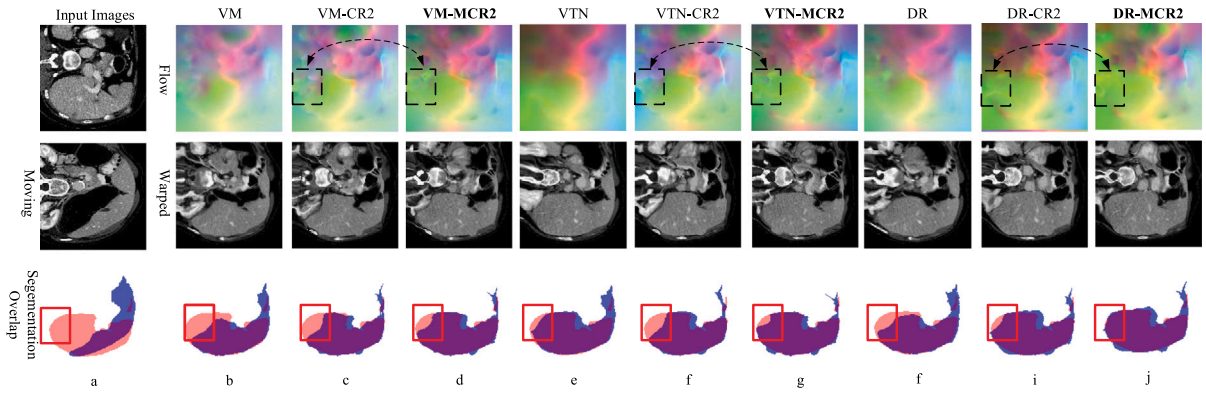

Fig. 11. Registration results from cascades with different strategies for a pair of images from the LiTS dataset.

with single scale loss or multi-scale loss in Eq. (13) whose results are presented in Table 6 where cases denoted with L in the brackets are the models trained with the multi-scale loss and other cases correspond to models trained with single scale loss. Here, the single scale loss is also computed according to Eq. (13) but only at l_4 layer of the last DR-Net in the cascade. As expected, the performance of cascaded models improves gradually when more sub-networks are incorporated irrespective the loss function used for training but the models trained with the multi-scale loss outperform those with the single scale loss. When the cascades are composed of two DR-Nets, the Dice Scores achieved by models trained with multi-scale loss are 0.4% and 0.4% higher than those achieved with models trained with the single scale loss on two test datasets. Simultaneously, the Similarity of the former models are 1.9% and 2.1% higher than those achieved with the latter models. Similarly, when three DR-Nets are used to construct the cascades, models trained with multi-scale loss also outperform the ones trained with the single scale loss where the Dice Scores of the former are 0.5% and 0.4% higher than those of the latter while the Similarities of the former are 1.6% and 1.2% higher than those of the latter on two test datasets. Pair-wise statistical t-tests show that all the above improvements are statistically significant ($p < 0.05$), which demonstrate that the introduction of the multi-scale loss along with our multi-scale cascading strategy can effectively enable the network learning to extract multi-scale structural information and hence significantly improve the registration accuracy of liver images without introducing any additional model parameters.

4.5. Comparisons with state-of-the-art cascaded models

For a comprehensive evaluation of the performance of our multi-scale cascaded model based on the DR-Net and some other state-of-the-art cascaded models reported in literature, both our model and the existing models are trained on the MSD and BFH datasets while evaluated on the Sliver and LiTS datasets. As before, the existing models are implemented with the open source codes provided by the authors under the same settings of hyper-parameters. Quantitative evaluation of the registration results from different methods is presented in Table 7. Here, the performance of the cascades VM-CR3 and IVTN-CR3 is investigated and compared with that of our DR-MCR2(L) and DR-MCR3(L). As can be seen from Table 7 that our DR-MCR2(L) with two sub-networks outperforms VM-CR3 and IVTN-CR3 with three sub-networks in terms of the Dice Scores on both test datasets, where the Dice Scores of our DR-MCR2(L) are 1.9% and 2.1% higher ($p < 0.05$) than those of the VM-CR3 while they are 0.8% and 1.1% ($p < 0.05$) higher than those of the IVTN-CR3 on both datasets. Although the Similarity measures achieved with our DR-MCR2(L) are a little bit lower than those with the VM-CR3 model on both datasets, they are 1.8% and 2.2% ($p < 0.05$) higher than those of the IVTN-CR3. When three DR-Nets are used to establish our cascades, the Dice Scores are 1.6% and 1.9% ($p < 0.05$) higher than those of the VM-CR3 while they are 1.5% and 1.7% ($p < 0.05$) higher than those of the IVTN-CR3. The Similarity measures achieved with our DR-MCR3(L) are 3.4% and 3.7% ($p < 0.05$) higher than those of the IVTN-CR3 but they are a little bit lower than those of the VM-CR3.

It can also be observed from Table 7 that the amount of parameters involved in our DR-MCR2(L) is only 59.6% of those involved in the IVTN-CR3 but our DR-MCR2(L) outperforms IVTN-CR3 in registration accuracy. Even if three DR-Nets are used to establish the cascade to achieve further improvement in registration accuracy, the amount of parameters of our DR-MCR3(L) is only 82.3% of those involved in the IVTN-CR3. Although our DR-MCR2(L) is larger than the VM-CR3, but the registration accuracy of the DR-MCR2(L) in terms of the Dice Scores is also better than that of the VM-CR3. All these imply that our method can achieve better registration results with a fairly lighter cascaded model involving fewer sub-networks.

Table 6
Performance ($\pm std$) of cascaded models trained with single and multi-scale loss.

Methods	Sliver		LiTS		Para. (Mb)
	Dice	Similarity	Dice	Similarity	
DR-MCR2	0.932 ± 0.022	0.833 ± 0.027	0.889 ± 0.049	0.828 ± 0.035	674
DR-MCR2(L)	0.936 ± 0.018	0.852 ± 0.022	0.893 ± 0.054	0.848 ± 0.031	674
DR-MCR3	0.938 ± 0.019	0.852 ± 0.024	0.895 ± 0.055	0.851 ± 0.031	931
DR-MCR3(L)	0.943 ± 0.018	0.868 ± 0.021	0.899 ± 0.056	0.863 ± 0.029	931

Table 7
Performance ($\pm std$) of different cascaded models.

Methods	Sliver		LiTS		Para. (Mb)
	Dice	Similarity	Dice	Similarity	
VM-CR3	0.927 ± 0.023	0.885 ± 0.019	0.880 ± 0.054	0.874 ± 0.028	172
IVTN-CR3	0.928 ± 0.023	0.834 ± 0.028	0.882 ± 0.050	0.826 ± 0.036	1131
DR-MCR2(L)	0.936 ± 0.018	0.852 ± 0.022	0.893 ± 0.054	0.848 ± 0.031	674
DR-MCR3(L)	0.943 ± 0.018	0.868 ± 0.021	0.899 ± 0.056	0.863 ± 0.029	931

For visually assessing the performance of different cascaded models, Fig. 12 gives the registration results from different models for an example image pair with large deformation from the Sliver dataset. It can be seen from the first row that the deformation fields predicted with our DR-MCR2(L) and DR-MCR3(L) are clearer and with higher contrast than those from the VM-CR3 and the IVTN-CR3, which implies that our multi-scale cascading strategy can fuse displace information in different scales and hence can depict better deformation fields. Inspecting the registration results from different models in the second row, we can see that the result from our two models look more similar to the fixed image than those from other models. The last row gives the extent of overlapping between liver regions in the fixed and warped moving images where the blue and red regions are the incoincident parts while the purple regions are the overlapped parts. Green and red boxes highlight the difference in results from different methods. Comparing the results in regions highlighted in red boxes we can see that the unsuccessfully registered regions in (d) and (e) of our two models are smaller than those in (b) and (c) of other two models. Regions marked with green boxes in (d) and (e) indicate that this local region is successfully registered with our two models while the corresponding part in (b) and (c) indicate failure registration with other two models. These visual assessments also demonstrate that our multi-scale cascading strategy and multi-scale loss can effectively enable the trained model better extracting multi-scale displacement details and hence better registering image details in large deformed image pairs.

5. Discussions and conclusions

In this work, we first propose a deformable registration network (DR-Net) for the registration of 3D liver CT images. Our DR-Net appears as a U-shaped model with some particular aspects: (1) A light weighted sequential Inception module is incorporated to enable the network integrating features of different scales; (2) A pyramidal input module is introduced so that multi-scale images can be input to the coding path, which makes the image features extracted with the network richer and more hierarchical; (3) An SCAM convolutional attention module is introduced to enable the network learning to simultaneously highlight useful channels or voxels while suppress irrelevant ones to our registration tasks; (4) An edge loss is introduced to enable the network better perceiving and hence better aligning textures of livers. We conduct extensive experiments on multi-center datasets to assess the effectiveness of the above mentioned modules on the registration accuracy whose results indicate that the registration accuracy improve gradually when the modules are incorporated in the network one after the other. Experimental results also demonstrate that when all the modules are incorporated in the network, our DR-Net outperforms state-of-the-art methods such as VoxelMorph and the IVTN.

Registration of medical image pairs with large deformation is a challenging problem. We have investigated the initial overlapping rates of labeled liver regions in between different image pairs in our four datasets. The results show that the initial Dice Scores of 37% image pairs in our datasets are lower than 0.7 while those of 66% image pairs are lower than 0.75. In some particular cases, the initial Dice Scores before registration can be as lower as 0.45. This poses urgent demands for developing registration methods suitable for registering largely deformed images. The difficulty in registering largely deformed image pairs lies in the simultaneously accounting for larger scale of deformation and smaller scale of fine details. The standard way to address this problem is to establish a cascaded network so that the large deformation can be decomposed to smaller ones and the registration can be implemented asymptotically. However, some problems remain unsolved for existing cascade methods. The first one is how to train the cascaded network and the other one is how to enable the network learning to extract multi-scale features for registering structures of different scales. To deal with these problems, we propose a multi-scale cooperative cascading strategy to establish cascaded network so that largely deformed images especially their detailed structures can be successively and accurately registered. This cascaded model integrates the deformation field information within and between sub-networks at different scales to synthesize the cascaded deformation fields with more detailed displacement information, thus reducing the difficulty in generating deformation field with large displacement so that more accurate registration of local small parts in largely deformed images can be achieved. To cooperatively train the cascaded network, not only the output of the final network layer but also the multi-scale outputs from different layers of the decoder in the last cascaded sub-network are used to calculate loss function. This strategy not only enables multi-scale constraints and better optimization of the cascaded network but also makes the loss function more compatible with the multi-scale outputs of the network and hence improving the registration accuracy. Experimental results on liver image datasets from different centers show that the multi-scale cascaded network can achieve cross-level superiority over other cascaded methods in registration accuracy and achieve better registration performance with fewer parameters.

It should be pointed out that the training datasets and the evaluation datasets of our model are from different medical centers, which demonstrates an excellency of our model in generalization. Hence, we can hope that our model can potentially be extended to all deformable image registration tasks for liver images from other centers. One possible limitation of this work would be the smoothness of the deformation field. Folding areas cannot be avoided in currently proposed cooperative cascading methods and may be amplified as the number of cascades increases, which brings challenges to the cascading network. Although we incorporate a regularization term in the loss for deformation fields, the folding in deformation fields still exists and the inverse consistency

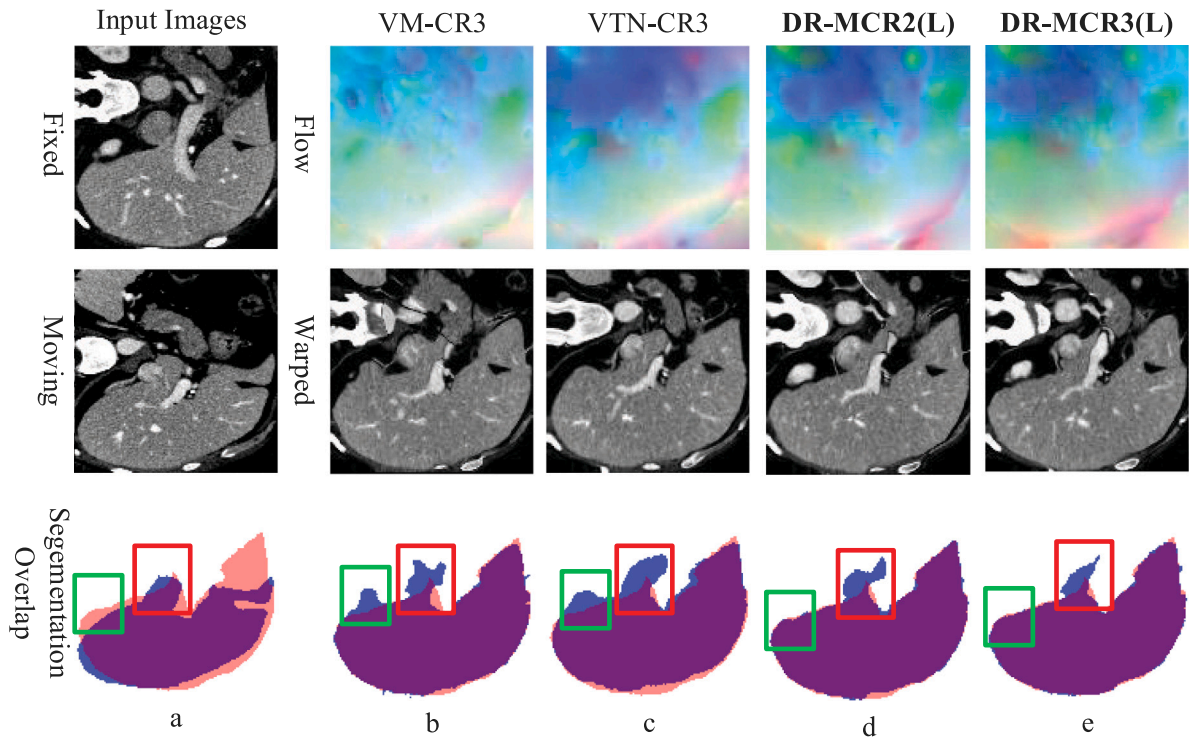

Fig. 12. Registration results from different cascades for a pair of images from the Sliver dataset.

cannot be guaranteed. Several approaches will be considered in our future work to solve this universal problem. These include ways such as taking a careful look at the regularization terms or using diffeomorphic deformation models for generating well-behaved deformation fields. Finally, if more hardware resources are available or a more efficient and lightweight cascade model can be exploited, we expect that the performance can be further improved by training with a bigger batch size. We will conduct a more in-depth exploration in the future work.

In conclusion, not only our proposed deep network (DR-Net) itself but also the cascade of them outperform the state-of-the-art methods and cascades in registration accuracy and model weight. Multi-scale input and multi-scale training strategies, cooperation between different layers of the network and cooperation between different sub-networks in the cascade all contribute to the superiority of our model. Excellent generalization performance of our model demonstrates that our model can be applied on datasets from other medical centers.

CRediT authorship contribution statement

Gangcheng Cai: Conceptualization, Methodology, Software, Writing – original draft. **Huaying Liu:** Writing – original draft, Writing – review & editing. **Wei Zou:** Formal analysis, Data curation. **Nan Hu:** Validation, Visualization. **JiaJun Wang:** Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The work reported in this paper was partly supported by the Natural Science Foundation of China No. 61473243 and the Natural Science Foundation of Jiangsu Province, China No. BK20171249.

References

[1] F.P. Oliveira, J.M.R. Tavares, Medical image registration: A review, Comput. Methods Biomech. Biomed. Eng. 17 (2) (2014) 73–93.
[2] P. Legg, P. Rosin, D. Marshall, J. Morgan, Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation, Comput. Med. Imaging Graph. 37 (7) (2013) 597–606.
[3] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C.M. Moore, M. Emberton, et al., Weakly-supervised convolutional neural networks for multimodal image registration, Med. Image Anal. 49 (2018) 1–13.
[4] G. Hermosillo, C. Chefd'Hotel, O. Faugeras, Variational methods for multimodal image matching, Int. J. Comput. Vis. 50 (3) (2002) 329–343.
[5] X. Huang, N. Paragios, D.N. Metaxas, Shape registration in implicit spaces using information theory and free form deformations, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1303–1318.
[6] B.B. Avants, N. Tustison, G. Song, et al., Advanced normalization tools (ANTS), Insight J. 2 (365) (2009) 1–35.
[7] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: A toolbox for intensity-based medical image registration, IEEE Trans. Med. Imaging 29 (1) (2009) 196–205.
[8] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2022) 3523–3542.
[9] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[10] A. Öztürk, A. Alhudhaif, K. Polat, Attention-based end-to-end CNN framework for content-based X-Ray image retrieval, Turk. J. Electr. Eng. Comput. Sci. 2021 (2021) 2680–2693.
[11] J. Fang, M. Zeng, X. Zhang, H. Liu, Y. Zhao, P. Zhang, H. Yang, J. Liu, H. Miao, Y. Hu, J. Liu, Deep metric learning with mirror attention and fine triplet loss for fundus image retrieval in ophthalmology, Biomed. Signal Process. Control 80 (2023) 104277.

[12] A. Öztürk, Image inpainting based compact hash code learning using modified U-net, in: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT, 2020, pp. 1–5.
[13] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, Voxelmorph: A learning framework for deformable medical image registration, IEEE Trans. Med. Imaging 38 (8) (2019) 1788–1800.
[14] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C.M. Moore, M. Emberton, S. Ourselin, J.A. Noble, D.C. Barratt, T. Vercauteren, Weakly-supervised convolutional neural networks for multimodal image registration, Med. Image Anal. 49 (2018) 1–13.
[15] H. Sokooti, B. De Vos, F. Berendsen, B.P. Lelieveldt, I. Išgum, M. Staring, Nonrigid image registration using multi-scale 3D convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 10433, 2017, pp. 232–239.
[16] S. Miao, Z.J. Wang, R. Liao, A CNN regression approach for real-time 2D/3D registration, IEEE Trans. Med. Imaging 35 (5) (2016) 1352–1363.
[17] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.
[19] S. Shan, W. Yan, X. Guo, E.I. Chang, Y. Fan, Y. Xu, et al., Unsupervised end-to-end learning for deformable medical image registration, 2017, arXiv preprint arXiv:1711.08608.
[20] D. Gu, G. Liu, X. Cao, Z. Xue, D. Shen, A consistent deep registration network with group data modeling, Comput. Med. Imaging Graph. 90 (2021) 101904.
[21] Y. Luo, W. Cao, Z. He, W. Zou, Z. He, Deformable adversarial registration network with multiple loss constraints, Comput. Med. Imaging Graph. 91 (2021) 101931.
[22] L. Qian, Q. Zhou, X. Cao, W. Shen, S. Suo, S. Ma, G. Qu, X. Gong, Y. Yan, J. Xu, L. Jiang, A cascade-network framework for integrated registration of liver DCE-MR images, Comput. Med. Imaging Graph. 89 (2021) 101887.
[23] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4040–4048.
[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[25] S. Abbasi, M. Tavakoli, H.R. Boveiri, M.A.M. Shirazi, R. Khayami, H. Khorasani, R. Javidan, A. Mehdizadeh, Medical image registration using unsupervised deep neural network: A scoping literature review, Biomed. Signal Process. Control 73 (2022) 1034444.
[26] B.D. de Vos, F.F. Berendsen, M.A. Viergever, H. Sokooti, M. Staring, I. Išgum, A deep learning framework for unsupervised affine and deformable image registration, Med. Image Anal. 52 (2019) 128–143.
[27] S. Zhao, T. Lau, J. Luo, I. Eric, C. Chang, Y. Xu, Unsupervised 3D end-to-end medical image registration with volume tweening network, IEEE J. Biomed. Health Inf. 24 (5) (2019) 1394–1404.
[28] S. Zhao, Y. Dong, E.I. Chang, Y. Xu, et al., Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10600–10610.
[29] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, 2015, CoRR abs/1506.02025, arXiv:1506.02025.
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2818–2826, http://dx.doi.org/10.1109/CVPR.2016.308.
[32] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023, http://dx.doi.org/10.1109/TPAMI.2019.2913372.
[33] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, P. Bourgeat, Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis, IEEE Trans. Med. Imaging 38 (7) (2019) 1750–1762, http://dx.doi.org/10.1109/TMI.2019.2895894.
[34] A.L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019, arXiv preprint arXiv:1902.09063.
[35] T. Heimann, van Ginneken, et al., Comparison and evaluation of methods for liver segmentation from CT datasets, IEEE Trans. Med. Imaging 28 (8) (2009) 1251–1265.