Deformable Image Registration with a Scale-adaptive Convolutional Neural Network

1st Yudi Sang
Dept. of Bioengineering & Dept. of Radiation Oncology
University of California, Los Angeles
Los Angeles, CA, USA
yudisang@ucla.edu

2nd Dan Ruan
Dept. of Radiation Oncology & Dept. of Bioengineering
University of California, Los Angeles
Los Angeles, CA, USA
druan@mednet.ucla.edu

***Abstract*—Multi-resolution hierarchical strategy is typically used in conventional optimization-based image registration to address large deformation and improve the chance of a good local minimum. A rough concept of the scale is captured in deep networks by the reception field of kernels, and it has been realized to be both desirable and challenging to capture convolutions of different scales simultaneously in registration networks. In this study, we propose an image registration network that is conscious of and self-adaptive to deformation of various scales. Dilated inception modules (DIMs) are proposed to incorporate receptive fields of different sizes in a computationally efficient way. Scale adaptive modules (SAMs) are proposed to guide and adjust shallow features using convolutional kernels with spatially adaptive dilation rate learned from deep features. DIMs and SAMs are integrated into the registration network which takes a U-net structure. The network is trained in an unsupervised setting and completes registration with a single evaluation run. Experiment with cardiac MRIs showed that the adaptive dilation rate in SAM corresponded well to the deformation scale. Evaluated with left ventricle segmentation, our method achieved a dice of (0.93±0.02), significantly better than SimpleElastix and networks without DIM or SAM. Assessment with respect to average surface distance was less than 2 millimeters (1.6 pixels), comparable to the best-performing SimpleElastix without statistical significance. Experiment with synthetic data also demonstrated the effectiveness of DIMs and SAMs, which leaded to a significant reduction in target registration error based on dense deformation field. The average registration time was 4 milliseconds for 2D image with size 256×256.**

***Index Terms*—deep learning, image registration, self-adaptive, scale, dilated convolution**

I. Introduction

Deformable image registration (DIR) is fundamental in medical image analysis: it aims to find the deformation vector field (DVF) to align fixed and moving image pairs. Typical applications include information fusion across various modalities or setups, longitudinal analysis, comparative studies across subjects, *etc*.

In conventional methods, deformable registration problems are posed in an optimization framework [1]–[3]. These methods usually seek a DVF to minimize an objective function consisting of image similarity and DVF regularity in an iterative process. Registration is often performed in multiple stages starting with affine registration, followed by coarse-to-fine scale DIR. This hierarchical multi-resolution strategy improves the methods' ability to handle large deformation and makes them less prone to undesirable local minima.

The development of deep learning methods provides an alternative to the iterative optimization methods [4]. A network can be trained to infer DVFs directly from image pairs, skipping the time-consuming iterative process. Training of the network can be supervised or unsupervised. The former learns from ground-truth DVF for each training image pair [5], [6], and the latter uses optimization formulations as in conventional methods and replaces the time-consuming iterative optimizers with deep networks [7]–[10].

Despite the benefit of efficiency, it is challenging for the deep learning methods to achieve highly physically and physiologically feasible registration since deformations of various scales need to be addressed by the network simultaneously. Tissue properties including elasticity and anisotropic discontinuity or sliding, vary across scales and spatial locations. This spatial heterogeneity in deformation scale is challenging for convolutional neural networks (CNNs), which typically use a single kernel size with a unique reception field at each level.

One way to address this problem is to implement the multi-resolution strategy by sequentially optimizing multiple networks with the warped image from the previous network passed to the subsequent network as the moving image input to progressively refine the DVF [7]–[9]. This increases the total number of network parameters and requires a prolonged training process. In addition, the repeated resampling of the moving image may accumulate errors and thus limit registration accuracy.

Another way to better accommodate the scale heterogeneity is by more sophisticated network architectures. U-net was proposed for segmentation tasks and was shown able to achieve good performance with relatively small datasets [11]. Its hierarchical structure and skip connections between encoder and decoder allow fusion of information across scales. The U-net structure has also been tested in registration as backbone networks in many studies [6], [9], [10]. Inception module was proposed with GoogleNet to capture multi-scale contextual features by using kernels with multiple sizes in parallel [12]. Several versions of the inception module have

This work was supported in part by a research grant from Varian Medical Systems, Inc.

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00096

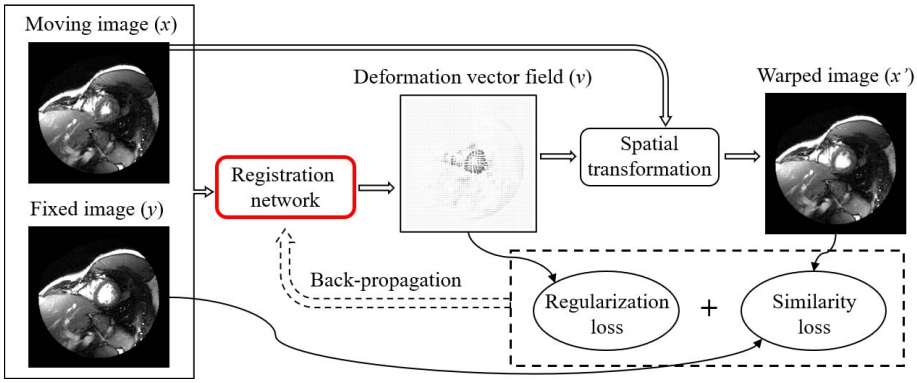

Fig. 1. Overview of the proposed registration framework.

been proposed in combination with factorized convolutions and residual blocks to improve the performance and efficiency [13], [14].

In this study, we propose a network for end-to-end deformable image registration that is conscious of and self-adaptive to information from various scales. Dilated inception modules (DIMs) are proposed to accommodate the need for larger reception fields and accommodate a wide range of scales efficiently. Scale adaptation modules (SAMs) are proposed to learn a spatial map of optimal scale based on dilated convolution and combine shallow and deep features in a self-adaptive setting.

II. Method

A. The Overall Image Registration Framework

As shown in Fig. 1, the proposed method consists of a registration network and a spatial transformation module. The registration network concatenates fixed and moving images as input and generates a DVF v, with which the moving image x is warped toward the fixed image y. Inside the spatial transformation module, a sampling grid is created using the input DVF. The input moving image x is sampled at these grid points to form the output warped image x' [15].

The loss function is defined as the weighted sum of the intensity match discrepancy and the regularity penalty:

$$L = L_s(y, x') + \lambda L_r(v), \tag{1}$$

where L_s is the image similarity loss, L_r is the DVF regularization loss, and λ is a balancing hyper-parameter. In this work, we use normalized cross-correlation (NCC) as the image similarity metric, but the method also applies to other choices such as mutual information [16].

$$\begin{aligned} L_s(y, x') &= 1 - \mathrm{NCC}(y, x') \\ &= 1 - \left\langle \frac{y - \bar{y}}{\|y - \bar{y}\|_2}, \frac{x' - \bar{x'}}{\|x' - \bar{x'}\|_2} \right\rangle. \end{aligned} \tag{2}$$

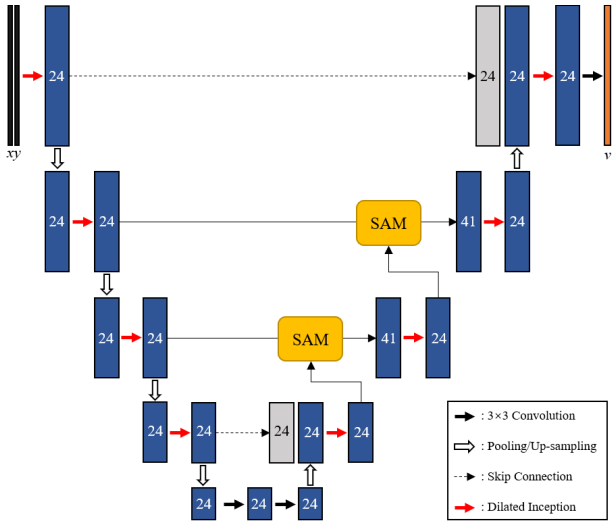

Fig. 2. Architecture of the registration network. Numbers inside the blocks indicate number of channels.

Bending energy penalty is used as the DVF regularization loss to penalize nonsmooth deformations and therefore encourage physical feasibility:

$$L_r(v) = \frac{1}{d_a d_b} \sum_{a=1}^{d_a} \sum_{b=1}^{d_b} \left[\left(\frac{\partial^2 v}{\partial a^2} \right)^2 + \left(\frac{\partial^2 v}{\partial b^2} \right)^2 + 2 \left(\frac{\partial^2 v}{\partial a \partial b} \right)^2 \right], \tag{3}$$

where a and b are spatial indices of the 2D image, and d_a and d_b are the corresponding spatial resolutions, respectively. During training, a back-propagation scheme is used to derive a DVF solution to minimize the objective in Eqn. (1).

B. Proposed Scale-adaptive Registration Network

The major contribution of this work is a scale-adaptive registration module. The module takes on a general U-net structure [11] to take advantage of the hierarchical structure and skip connections for effective learning of features at all scales. Scale adaptation is achieved with the introduction of DIMs and SAMs.

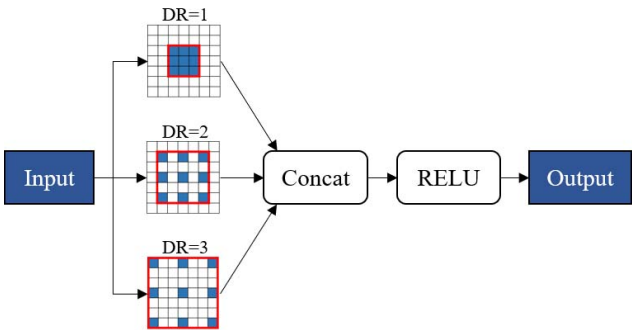

Fig. 3. Dilated inception module.

Fig. 2 shows the architecture of the networks. In both of the encoding and decoding paths, most of the standard 3×3 convolution layers in the conventional U-net are replaced with DIMs to increase the receptive field size (details in Sec. II-C).

In addition, SAMs are introduced into the skip connecting paths at the second $(1/2)$ and third $(1/4)$ resolution levels to better extract shallow features using kernels with adaptive receptive fields learned from deep features (details in Sec. II-D).

All the standard 3×3 convolution layers use a stride of 1, zero-padding and ReLU activation, except the last layer, which uses linear activation. Average pooling and bilinear interpolation with a scaling factor of 2 are used in the down- and up-sampling layers, respectively.

C. Dilated Inception Module (DIM)

Dilated convolution, also know as atrous convolution, enlarges the receptive field by inserting spacing into convolutional kernel according to the dilation rate (DR) [17], [18]. A dilated convolution operation can be written as:

$$g[a,b] = \sum_{m=1}^{M} \sum_{n=1}^{N} f\left[a + r \cdot m, b + r \cdot n\right] \cdot h[m,n], \quad (4)$$

where f is the input feature map, g is the output feature map, h is the convolutional kernel with effective kernel size $M \times N$, and r is the dilation rate. m and n, and a and b are the spatial indices for the convolutional kernal and feature map, respectively. When $r = 1$, the operation reduces to a standard convolution layer. Compared to standard convolution with large kernel size, dilated convolution is able to achieve large receptive fields without additional network parameters. Compared to using large (> 1) stride in a convolution layer, dilated convolution keeps the spatial resolution and may preserve spatial information better.

Inception blocks have been used for classification and detection tasks in computer vision and were shown to improve performance with modest increase in computational cost [12]. An inception block combines convolutional operations with multiple kernels of different sizes, and concatenate their output features to form a single output to be fed into the next layer. The output of inception block contains features with effective receptive fields of different sizes.

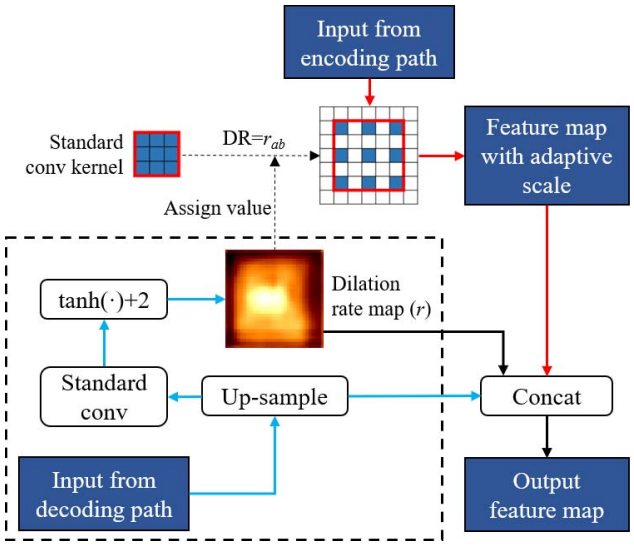

Fig. 4. Scale adaptation module. "Input from decoding path" is from the lower resolution level, and its path within the module is indicated by the blue arrows. "Input from decoding path" is via the skip connection in the U-net, and its path within the module is indicated by the red arrows.

We propose to integrate dilated convolution kernels into an inception block to combine the benefit of computational efficiency in dilated convolution and the richness in scale of the learned features from inception setup. We refer to this design as a dilated inception module (DIM). As shown in Fig. 3, three 3×3 dilated convolutions (*i.e.*, $M = N = 3$) with dilation rates $r = 1, 2, 3$ are apply to the input feature map in parallel to extract features with different receptive fields. The convolution results are then concatenated and activated with a ReLU function to form the output. In this work, we use 8 channels for each of the 3 kernels, resulting in an output feature map with 24 channels.

D. Scale Adaptation Module (SAM)

In conventional U-net, shallow features in the encoding path are passed to the decoding path through skip connections. Here we proposed to further utilize the contextual information provided by the deep features to guide the propagation of shallow features in the skip connections according to local scales. We propose a scale adaptation module (SAM) to integrate the features in a self-adaptive way.

As shown inside the dashed box in Fig. 4, the input feature map from the previous resolution level in the decoding path is first up-sampled by 2 and passed through a standard convolution layer. It is then mapped to the range of $(1, 3)$ with tanh(·)+2 as an activation function to form the corresponding estimate of dilation rate map (DRM). Each pixel value in the DRM is used to assign the dilation rate of a 3×3 dilated convolutional kernel at each corresponding spatial location. This convolution with spatially adaptive dilation rate is applied to the input features from encoding path to generate a feature map with adaptive scale. Finally, this feature map is concatenated with the up-sampled decoding features and the

DRM to form the output of SAM. The up-sampled decoding features, the adaptive dilated convolution, and the DRM have 24, 16, and 1 channels, respectively, yielding an overall output feature map of 41 channels (as illustrated in Fig. 2). The standard convolution layer for DRM generation uses kernel size of 3×3 for the SAM at the $1/2$ level and 5×5 at the $1/4$ level.

The implementation of the adaptive dilated convolution is slightly different from Eqn. (4). Since the dilation rate is limited within the range $(1, 3)$, we first compute the feature maps from dilated convolutions with $r = 1, 2, 3$, respectively. Note that the parameters for the three dilated convolutional kernels are shared because they arise from the same standard kernel. For fractional dilation rates, linear interpolation of the three feature maps at each spatial location is used as a surrogate:

$$g(r) = t_1 \cdot g(1) + t_2 \cdot g(2) + t_3 \cdot g(3), \tag{5}$$

where $t_1 + t_2 + t_3 = 1$, numbers inside the parentheses indicate dilation rates.

Our method is faster and requires less memory than the approach in [19], where the bilinear interpolation of the feature map involves a large interpolation array, whose high dimensionality makes it computationally inefficient.

III. Experiments and Results

The network was implemented using TensorFlow. It was tested on cardiac MRIs with left ventricle segmentations for quantitative evaluation, and on synthetic image pairs with dense ground-truth DVFs. The proposed method was compared against a conventional B-spline-based SimpleElastix method [3] and three simplified versions of the registration network: U-net alone, U-net with DIM, and U-net with SAM. The objective function was kept the same for SimpleElastix and the networks. In SimpleElastix, multi-resolution strategy was used, with 20 optimization iterations in each of the 4 resolution levels. The experiments were performed on an NVIDIA GTX 1080 Ti GPU and an Intel i7-6700HQ 3.5 GHz CPU.

A. Experiments on Cardiac MRIs

1) Data: The method was tested on 2D cardiac MRI sequences obtained from Sunnybrook Cardiac Data [20], which contains 45 4D short-axis cardiac cine MR scans, each containing 20 frames that cover the cardiac cycle. The image resolution was 256×256, with 10 slices, pixel spacing 1.25 mm, and slice thickness 8 mm. Segmentations of left ventricles were provided in the dataset at end-diastole (ED) and end-systole (ES) frames. 45 4D scans were divided into training, validation, and testing sets, containing 30, 5, and 10 3D videos, respectively. Fixed and moving image pairs were prepared by picking 2D slices from the same 4D scan, at the same slice position but at different time points in the cardiac cycle. Down-sampled in the temporal domain, 27,000 2D image pairs were used for training eventually. A typical image pair is shown in Fig. 5(a,b).

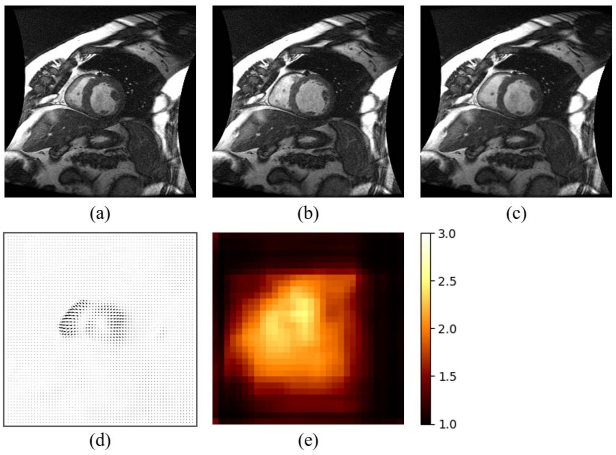

Fig. 5. Example registration result in our method. (a) Moving image. (b) Fixed image. (c) Warped image. (d) DVF. (e) DRM at $1/4$ level.

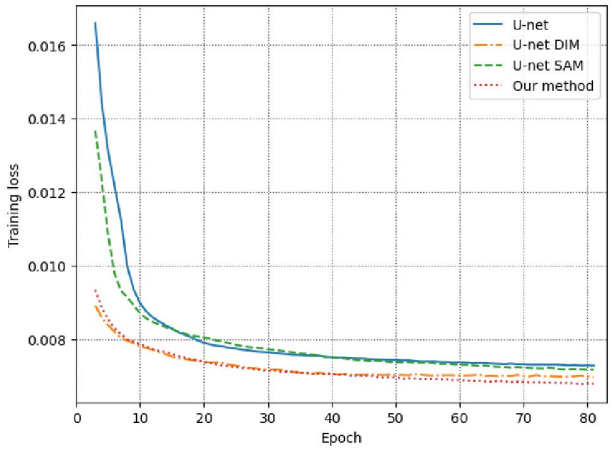

Fig. 6. Learning curves showing the total loss during the training of the networks. Values for the first two epochs are not displayed.

2) Network Training: Each registration network was trained in mini-batches of 8 image pairs for 80 epochs. The balancing weight λ in Eqn. (1) was set to 500. ADAM optimizer with learning rate of 10^{-4} was used.

3) Evaluation: Using the provided segmentations of left ventricles, we computed the following metrics to evaluate the registration performance.

Dice coefficient between the propagated segmentation mask $M_{x'}$ and the segmentation masks on the fixed image M_y:

$$Dice = \frac{2|M_{x'} \cap M_y|}{|M_{x'}| + |M_y|}. \tag{6}$$

Average surface distance (ASD) between the propagated and fixed segmentation contours:

$$ASD = \frac{\sum_{p_{x'} \in C_{x'}} dist(p_{x'}, C_y) + \sum_{p_y \in C_y} dist(p_y, C_{x'})}{|C_{x'}| + |C_y|}, \tag{7}$$

TABLE I
RESULTS OF THE CARDIAC MRI REGISTRATION EXPERIMENT. RESULTS ARE PROVIDED AS MEAN ± STANDARD DEVIATION (Z-VALUE FROM WILCOXON SIGNED-RANK TEST). Z-VALUES THAT INDICATE STATISTICAL SIGNIFICANCE ARE UNDERLINED.

	Dice	ASD (mm)	Time (s)
SimpleElastix	0.92±0.03 (2.13)	**1.65±0.98** (1.89)	5.63
U-net	0.91±0.04 (4.72)	1.73±1.37 (3.11)	**0.002**
U-net DIM	0.93±0.02 (2.00)	1.69±1.01 (1.60)	**0.002**
U-net SAM	0.91±0.03 (3.56)	1.73±1.33 (2.95)	0.004
Our method	**0.93±0.02**	1.68±0.95	0.004

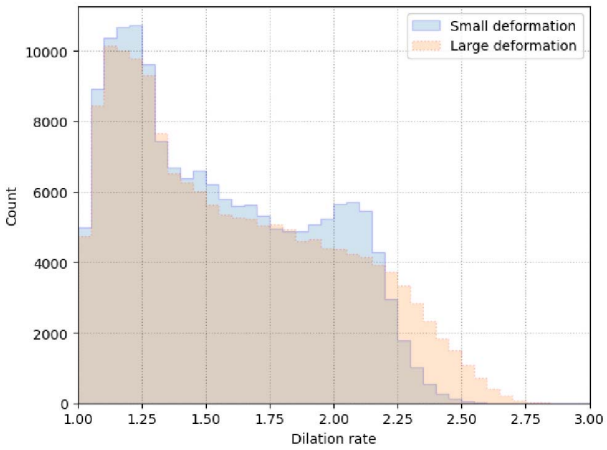

Fig. 7. Histograms showing the distributions of dilation rates in the SAM for testing image pairs with small and large deformations. The number of pixels are the same, and the total counts are equal for the two input settings.

where $p_{x'}$ and p_y are points on the propagated contour $C_{x'}$ and the contour on the fixed image C_y, respectively.

4) Results: Fig. 6 shows the learning curves of the networks. It can be observed that both DIM and SAM effectively reduced the final loss value and DIM provided more significant improvement. The self-adaptation mechanism in SAM made the networks slightly slower to converge. At around the 40th epoch, SAMs began to surpass the counterparts without SAM.

Fig. 5 shows an example registration result. It can be seen that large dilation rate values in the DRM in (e) correspond to the spatial locations with large deformation in (d).

Table. I shows the quantitative results. Wilcoxon signed-rank tests were performed to examine the statistical significance (at significance level of 0.05, Z_{crit}=1.96). In terms of dice coefficient, our method achieved the best result among all the methods tested, with statistical significance. In terms of ASD, our method outperformed other networks and was close enough to the best-performing SimpleElastix results without statistical significance. Introduction of the SAM increased the network registration time by 2 milliseconds, but the network still offers more than 3 orders of magnitude speed-up over SimpleElastix.

To further appreciate the performance of the SAM, we

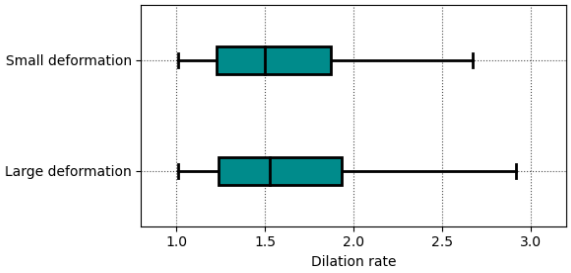

Fig. 8. Boxplot showing the quantiles of dilation rates for small and large deformations.

TABLE II
STATISTICS OF THE DILATION RATE FOR LARGE AND SMALL DEFORMATIONS IN THE TESTING SET.

	Proportion	Percentile		
	$r > 2.25$	75th	85th	95th
Small deformation	0.023	1.87	2.02	2.18
Large deformation	0.008	1.93	2.12	2.35

calculated histograms of the dilation rates in the SAM at the $1/4$ resolution level. A comparison was made between DRMs for image pairs of (a) adjacent frames near ED with small deformations and (b) ED and ES frames with large deformations, as shown in Fig. 7. It can be observed that the distribution of the estimated dilation rates is in agreement with the qualitative motion levels: with the larger deformations between ED and ES taking a larger proportion of high dilation rate. Fig. 8 shows the quantiles of the dilation rate distributions. Table. II shows the proportion of dilation rate values larger than 2.25 $P(r > 2.25)$ and the percentiles of the dilation rates.

B. Experiments on Synthetic Data

1) Data Generation: Given the absence of ground-truth DVFs in clinical images, we synthesized DVFs and images as digital phantoms for a quantitative evaluation on dense DVFs. Data generation was based on the same dataset as described in Sec. III-A1. As shown in Fig. 9, first we used SimpleElastix (with the same objective function and setup) to generate DVFs between ES and ED frames of the images. These DVFs were multiplied by a scaling factor α uniformly distributed within the range $[0.5, 1.2]$ for data augmentation. The scaled DVFs were taken as ground truths and were used to generate the warped images in the forward direction with corresponding moving images. Training, validation, and testing sets contained 4800, 800, and 1600 2D synthetic samples, respectively.

2) Network Training: Each registration network was trained in mini-batches of 8 image pairs for 500 epochs. The balancing weight λ in Eqn. (1) was set to 500. ADAM optimizer with learning rate of 10^{-4} was used.

3) Evaluation: Root mean squared errors (RMSEs) to the ground-truth DVFs were calculated as a dense version of target

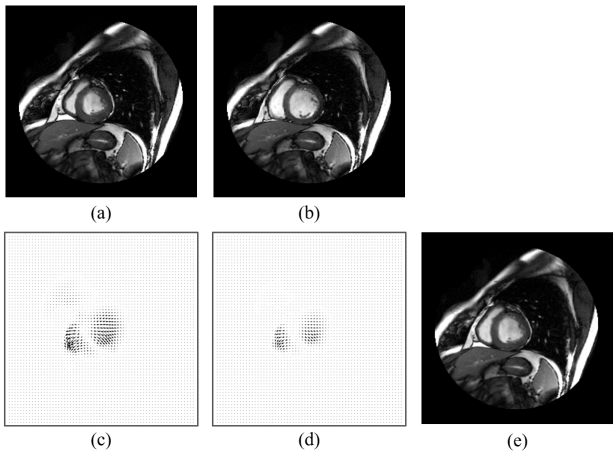

Fig. 9. Example synthetic data. (a) Moving image. (b) Fixed image in SimpleElastix. (c) SimpleElastix DVF. (d) Scaled DVF with $\alpha = 0.6$. (e) Forward generated fixed image.

TABLE III
RESULTS OF THE SYNTHETIC IMAGE REGISTRATION EXPERIMENT. RESULTS ARE PROVIDED AS MEAN ± STANDARD DEVIATION AND p-VALUE FROM PAIRED T-TEST.

	TRE (mm)	p-value
U-net	0.024±0.020	2.75×10^{-8}
U-net DIM	0.019±0.011	9.53×10^{-3}
U-net SAM	0.022±0.020	2.12×10^{-7}
Our method	**0.017±0.009**	-

registration error (TRE) to indicate registration performance:

$$TRE = \sqrt{\frac{1}{d_a d_b}\sum_{a=1}^{d_a}\sum_{b=1}^{d_b}\|v^*_{ab} - v_{ab}\|^2}, \tag{8}$$

where v^* and v are the ground-truth and result DVFs.

4) Result: Paired t-tests were performed to examine the statistical significance. As shown in Table. III, our method achieved the lowest TRE among all the networks, with statistical significance.

IV. DISCUSSION AND CONCLUSION

In this study, we have presented a deep neural network for deformable image registration. The major contribution of this work is the introduction and integration of DIMs and SAMs, which address the heterogeneous scale problem in an efficient and self-adaptive way, as opposed to CNN architectures using a fixed kernel size in each convolution layer, which has to be designated a priori. The DIMs explicitly enlarge the effective receptive field without additional network parameters. The SAMs process the shallow features from the encoding path through skip connections guided by deep decoding features and allow network parameters to be updated based on the spatial region that covers the local deformation scale. The effectiveness of the modules was shown in the experiments. Our method achieved better or comparable results compared to SimpleElastix, where the scale heterogeneity was addressed with a multi-resolution strategy.

An example DRM at the $1/4$ level was presented in Fig. 5. However, in all of our experiments, DRM at the $1/2$ level converged to 1 (*i.e.*, standard convolution), regardless of the kernel initialization methods, presence or removal of the DIMs, *etc*. One possible explanation was that features at the $1/2$ level were close to the output layer and focused more on the local scale for DVF refinement. We plan to further investigate this problem and determine the optimal SAM placement.

In SAM, we used linear interpolation of pre-computed feature maps with integer dilation rates as computationally efficient surrogates to convolutional kernels with fractional dilation rates. This simplified method was faster and required much less memory than the bilinear interpolation method in [19], which is important in a 3D setup. The 2D pilot experiments performed in this study did not properly account for out-of-plane tissue movement, and we are currently extending the scope to 3D space and settings.

REFERENCES

[1] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: a survey," *IEEE Trans. Biomed. Eng.*, vol. 32, no. 7, p. 1153, 2013.
[2] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Comput. Methods. Biomech. Biomed. Engin.*, vol. 17, no. 2, pp. 73–93, 2014.
[3] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A user-friendly, multi-lingual library for medical image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 134–142.
[4] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Mach. Vision Appl.*, vol. 31, no. 1, p. 8, 2020.
[5] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.
[6] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 266–274.
[7] C. Shu, X. Chen, Q. Xie, and H. Han, "An unsupervised network for fast microscopic image registration," in *Proc. SPIE Med. Imag.*, vol. 10581, 2018, p. 105811D.
[8] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, 2019.
[9] S. Zhao, Y. Dong, E. I. Chang, Y. Xu *et al.*, "Recursive cascaded networks for unsupervised medical image registration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10 600–10 610.
[10] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, 2019.
[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.
[15] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[16] B. D. de Vos, B. H. van der Velden, J. Sander, K. G. Gilhuijs, M. Staring, and I. Išgum, "Mutual information for unsupervised deep learning image registration," in *Proc. SPIE Med. Imag.*, vol. 11313, 2020, p. 113130R.
[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv:1511.07122*, 2015.
[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.
[19] X. Guo, Z. Chen, and Y. Yuan, "Complementary network with adaptive receptive fields for melanoma segmentation," in *Proc. IEEE 17th Int. Symp. Biomed. Imaging*, 2020, pp. 2010–2013.
[20] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac MRI," *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 49, 2009.