2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Image-Based 3D Ultrasound Reconstruction with Optical Flow via Pyramid Warping Network

Yanting Xie, Hongen Liao, Daoqiang Zhang, Lei Zhou, Fang Chen*

Abstract— **3D Ultrasound (US) contains rich spatial information which is helpful for medical diagnosis. However, current reconstruction methods with tracking devices are not suitable for clinical application. The sensorless freehand methods reconstruct based on US images which is less accuracy. In this paper, we proposed a network which reconstructs the US volume based on US images features and optical flow features. We proposed the pyramid warping layer which merges the image features and optical flow features with warping operation. To fuse the warped features of different scales in different pyramid levels, we adopted the fusion module using the attention mechanism. Meanwhile, we adopted the channel attention and spatial attention to our network. Our method was evaluated in 100 freehand US sweeps of human forearms which exhibits the efficient performance on volume reconstruction compared with other methods.**

I. Introduction

Ultrasound (US) imaging is a commonly used diagnosis imaging technology in various clinical applications, which combines several advantages such as real time, safe and low cost. But 2D US image has the disadvantage that the information is not intuitive enough. Compared to traditional 2D US image, 3D US contains richer spatial information which is often highly desired [1]. Thus, reconstructing 3D US volume from a sequence of 2D US images has become one of the important technologies.

Traditional methods used in 3D reconstruction usually combine the US probes with tracking devices. Based on the six degrees of freedom (6Dof) motions of all frames related to the first one obtained by the tracking devices, 3D volume is reconstructed from a sequence of frames. Common tracking devices contain mechanical tracking systems, optical tracking systems and electromagnetic tracking systems [2,3], These methods with tracking devices have the disadvantages such as not flexibility, restricted in usage scenarios and susceptible to interference, which are not suitable for clinical application.

Without using tracking devices, some methods are proposed to reconstruct the volume only based on features extracted from the US images. Speckle decorrelation is one of the most common traditional sensorless freehand methods which estimates the relative transformation based on the speckle features of neighboring US images [4]. Meanwhile, deep learning methods based on convolutional neural networks (CNN), which are widely used in extracting images features, are used in US images reconstruction. In recent years, some research attempts to directly estimate the inter-frame motion between neighboring US images using CNN which performs better than traditional methods [5,6,7].

As the motion information of each pixel between neighboring images, optical flow is the important information which is frequently used in motion estimation. In US reconstruction, optical flow is commonly used to estimate the in-plane motion. What is more, some research tries to append the optical flow features to CNN. But these studies just simply concatenate the optical flow features extracted from optical flow extractors [8] or stack the original image and optical flow as input data [9], which cannot make full use of optical flow information.

In this paper, we proposed a network which combines the optical flow features with warping operation. The operation warps the feature map of second image toward first one based on optical flow. What is more, we obtain different scales feature maps of US images and optical flow information while low-resolution levels exploit low-frequency B-mode information [8]. Thus, we define a pyramid warping layer which warping the optical flow information in pyramid levels. To fuse the different scales warped features of each pyramid level, we also use the fusion module based on attention mechanism. What is more, we use attention modules which focus on the interested regions of the feature maps. Motivated by the convolutional block attention module proposed in [9], we append the channel attention and spatial attention to our network.

II. Method

Our method is proposed to reconstruct US volume only from image sequences. Fig. 1 depicts the architecture of our reconstruction network. In this section, we firstly introduce the architecture of our network. Then, we introduce the implementation of pyramid warping layer and attention modules used in the network. Finally, we introduce the loss function used to improve the training process for the whole network.

A. Overall Network Structure

The inputs of the network are two adjacent US images, and the size of each frame is 256×256. The network consists of two pathways which process the original US images and warped

*Research supported by Foundation. National Nature Science Foundation of China grants (U20A20389, 61901214), China Postdoctoral Science Foundation (2021T140322, 2020M671484), Jiangsu Planned Projects for Postdoctoral Research Funds(2020Z024), High-level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China

F. Chen, Y. Xie, and D. Zhang are with the Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 210016, China. F. Chen is the corresponding authors of this paper (E-mail: chenfang@nuaa.edu.cn)

L. Zhou and H. Liao are with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 10084, China.

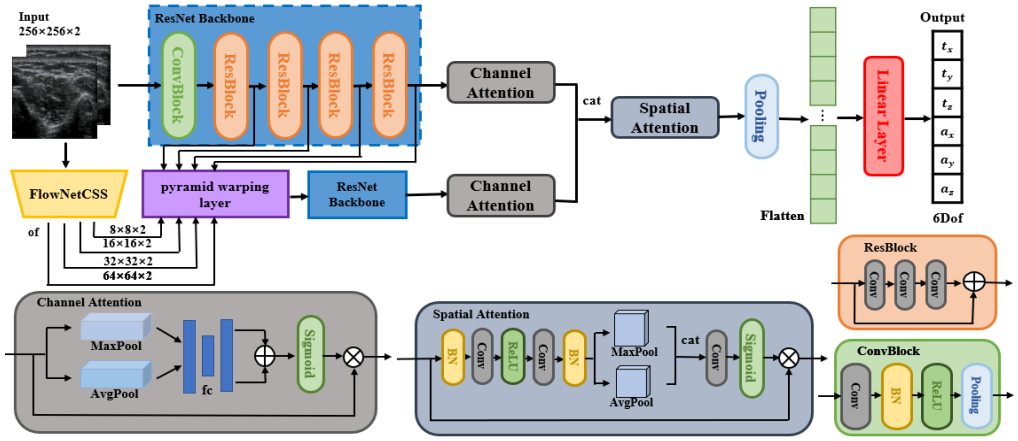

Fig. 1. Overview of the proposed network architecture

features based on optical flow information separately. We adopt ResNet backbone as the feature extractor which consists of a Convblock and four residual blocks. In the original US images pathway, we just append a channel attention after ResNet backbone.

In the warped features pathway, we use FlowNetCSS module and pyramid warping layer to compute the warped features based on optical flow information. The warped features are calculated by a decoder with pyramid architectures which fuses the optical flow features from different scales. And the shape of the outputs in pyramid warping layer is 256×256×2. Then, same as the original US images pathway, we use the ResNet backbone and channel attention to compute the feature map of warped features pathway.

Both pathways output a feature map which is 8×8×512. After concatenating these two feature maps, a spatial attention is followed. Then, we append a global average pooling (GAP) which reduces the features to a 1024-dimensional feature vector. Finally, the last two fully connected layers are used to reduce the feature vector to a 6-dimensional vector. The outputs of the network are 6 parameters consisting of 3 translations parameters and 3 rotation angles.

B. FlowNetCSS Module

The FlowNetCSS module is part of FlowNet2.0 proposed in [10]. It is stacked by FlowNetC and FlowNetS [11], which use FlowNetC as the first block, followed by two FlowNetS blocks. We train the module with a fine-tuning schedule. The module of our network loads the weights trained by the Chairs→Things3D schedule from [10] and then fine-tuning on our US images dataset. The inputs of the FlowNetCSS module are two adjacent US images, and the size of each frame is 256×256. The outputs are predicted optical flow features with 4 different scales (64×64×2, 32×32×2, 16×16×2, 8×8×2), which contains the motions information of each pixel in x dimension and y dimension.

C. Pyramid Warping Layer

Warping operation is widely used in optical flow prediction network, which warps the second image to the first one based on the predicted optical flow information. To reduce the number of parameters, [12] proposed to warp the features of the moved images instead of original images. Based on the coarse-to-fine scheme, we use the pyramid approach to combine the feature maps of different scales. Thus, we define a pyramid warping layer which warps the features in different scales and merges with the pyramid scheme. Fig. 2 depicts the architecture of pyramid warping layer.

The inputs of pyramid warping layer consist of the outputs of FlowNetCSS module and the outputs of each residual block of the ResNet backbone. Both of the outputs contain 4 feature maps of different scales (64×64×64, 32×32×128, 16×16×256, 8×8×512). Warping operation is defined in [12], which estimates the preliminary motion in the second image based on the intermediate optical flow information. Before warping operation, we use a 1×1 convolution layer to reduce the number of channels to 64. After normalizing the feature maps, we warp the 4 scales of US feature maps with the optical flow features with the same scales. Thus, we get 4 different scales of warped feature maps.

To merge different scales of warped feature maps, we use upsample operation which is commonly used in decoder. Starting with the low-resolution features exploited from low-frequency US images, the upsampled low-scaled feature maps are merged with the high-scaled feature maps based on fusion module using the attention mechanism [13]. The concatenated feature maps of F_A and F_B passes through a series of convolutions to obtain a feature map F_{AB} with 2 channels. Then, the weight maps of two scale features S_A and S_B are calculated using SoftMax function for each channel of F_{AB}. Thus, the merged feature F_{fusion} is obtained as a weighted sum:

$$F_{fusion} = S_A \times F_A + S_B \times F_B. \quad (1)$$

where the element-wise operation ($\times$) are performed between weight maps S and two scale features F to get the fused feature map F_{fusion}.

After fusing all warped features, we upsample the merged feature map to the original US images size which is 256×256. At the final layer, two 1×1 convolution layers followed by ReLU are used to reduce the number of channels to 2.

D. Channel Attention and Spatial Attention

To focus on interested regions of the feature maps, we add the attention mechanism to the network, which includes channel attention and spatial attention. These two attention modules separately emphasize the features along two principal dimensions: channel and spatial axes [9]. Fig. 1 depicts the architectures of channel attention and spatial attention.

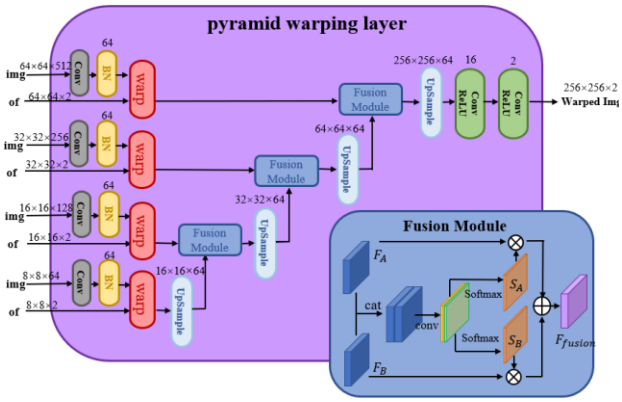

Fig. 2. The architecture of the pyramid warping layer

We adopt the channel attentions after ResNet backbones in two pathways of the network, which weights the features among the channels. The max pooling operation and average pooling operation are used to obtain the max-pooled and average-pooled features simultaneously. Then, the outputs of two pooling operations forward to the weight-shared multi-layer perceptron (MLP). We merge the features calculated from average-pooled features and max-pooled features. After the sigmoid activation function, we get the weights of the features of each channel. In short, the channel attention is computed as:

$$M_c(F) = F \times \delta\left(W\left(AvgPool(F)\right) + W\left(MaxPool(F)\right)\right) \quad (2)$$

where δ denotes the sigmoid function, W are the weights of the MLP.

The spatial attention is adopted after the concatenation of two pathways outputs, which emphasizes the features of each spatial feature map. Same as the channel attention, we simultaneously use max pooling operation and average pooling operation to calculate two 2D maps. Before the pooling operation, we adopt two standard convolution layers to reduce the number of parameters of feature maps. After the pooling operation, we merge the max-pooled map and average-pooled map with concatenate operation. With a standard convolution layer which reduces the number of channels to 1, we get a 2D spatial attention map which weights the spatial information of each feature map. The spatial attention is computed as:

$$M_S(F) = F \times \delta\big(f([AvgPool(F); MaxPool(F)])\big) \quad (3)$$

where δ denotes the sigmoid function, f represents the standard convolution operation with the filter size of 1×1.

E. Loss Function

The loss function of the proposed network consists of two components, which are mean absolute error loss ($Loss_{MAE}$) and transformation matrix loss ($Loss_{Tran}$). The MAE loss is used to estimate the distance between six degrees of freedom (6Dof) and ground truth. This loss function is shown as equation (4).

$$Loss_{MAE} = \frac{1}{6}\sum_{i=1}^{6}\left|\theta_i^{Out} - \theta_i^{GT}\right| \quad (4)$$

where θ_i^{Out} and θ_i^{GT} are the parameters of 6Dof of output and ground truth. We also define a loss function to estimate the motion between two US image frames. This loss function is calculated by computing the Frobenius norm between 4×4 transformation matrixes converted from 6Dof of output and ground truth. With this loss function, we can measure the motions of adjacent frames based on translation and rotation information. The loss function is based on equation (5).

$$Loss_{Tran} = Mean\left(\sum\left\|T_{GT}T_{Out}^{-1} - I\right\|_F\right) \quad (5)$$

where T_{GT} and T_{Out} are the transformation matrixes computed by 6Dof of output and ground truth, I is the 4×4 identity matrix.

III. EXPERIMENTS AND RESULTS

A. Experiment Dataset and Metric

Our dataset has 100 freehand US sweeps on human forearms which totally contains 19502 frames B-mode US images. All US sweeps are acquired on a Mindray DC 6E II US machine. For each frame of US images, we use the NDI Polaris Vicra to record the US probe positions which fixed an optical marker on the probe. The spatial and temporal calibrations of the probe are implemented by Plus Toolkit [14].

We use 10-fold cross validation scheme to evaluate the performance of models. During the experiments, we mainly use mean distance error and final drift error to evaluate the performance of reconstruction. The mean distance error is the mean distance of each frame between ground truth and predict results in one sweep, and the final drift error only calculate the distance of the final frame.

B. Reconstruction Performance and Discussion

To evaluate the effectiveness of the network our proposed for 3D US volume reconstruction, in this subsection, we compared it with other methods. Table 1 summarizes the reconstruction results of each method based on mean distance error and final drift error.

The approach of "Linear" means that we first calculate the mean 6Dof vector of the training set and then apply this fixed vector to the testing cases. The approach of "2D CNN" is the method proposed in [7] which only uses a series of 2D convolutions. The approach of "DCL-Net" is the method proposed in [5] which using 3D convolutions taking multiple adjacent frames. "Without OF" refers to the network our proposed without optical flow pathway which reconstructs the volume with image features. "Without PWL" refers to the network our proposed without pyramid warping layer, which concatenates the US images and optical flow features in channel dimension as input. "Without AM" refers to the network our proposed without channel attention and spatial attention.

As can be seen from experimental results, the method our proposed outperforms all other methods in 3D Ultrasound images reconstruction, which has the lowest values of mean distance error and final drift error. Depending on the results of first four lines in Table 1, our method has the better reconstruction results compared with the methods proposed in recent years. Compared with DCL-Net which was proposed in [5] in 2020, although our proposed method has the similar reconstruction performance in simple cases which has the similar values in minimize value of errors, our method performs better in difficult cases which has the lower values in maximize value of errors and average value of errors.

TABLE I. EXPERIMENTAL RESULTS OF DIFFERENT RECONSTRUCTION METHODS

Methods	Distance Error (mm)			Final Drift (mm)		
	min	*max*	*avg*	*min*	*max*	*avg*
Linear	4.66	14.57	7.86	6.77	26.69	14.65
2D CNN	3.13	12.88	7.64	4.15	23.23	12.28
DCL-Net	1.48	12.01	5.29	3.05	22.10	9.74
Proposed	**1.35**	**10.52**	**4.73**	**3.03**	**16.82**	**8.55**
Without OF	2.03	11.70	5.73	3.36	21.34	10.51
Without PWL	1.91	10.84	5.44	4.04	20.09	10.20
Without AM	1.79	11.67	5.49	3.46	21.88	10.29

What is more, we validate the modules used in our network based on the results of last four lines in table 1. Compared with the results of "Proposed" and "Without OF", we validate that the optical flow features which contain the motion information of images are helpful for reconstruction. And compared with the results of "Proposed" and "Without PWL" which means the network our proposed only without pyramid warping layer, we validate that the pyramid warping layer our proposed is helpful for US reconstruction which makes use of different scales optical flow features effectively. Meanwhile, compared with the results of "Proposed" and "Without AM" which means the network our proposed without attention modules, we validate that the channel attention and spatial attention we adopted will promote the results of reconstruction.

To show the effect of reconstruction, we visualize the 3D trajectories reconstructed of cases in testing dataset in Fig. 3, one good case, one poor case and one median case. As can be seen form the Fig. 3, the predicted reconstructions (green line) only severely deviate in the poor case which final drift error is 16.82 mm. What is more, our predicted reconstructions have a smoother trajectory which reduce the interference of noise produced by jittering.

IV. CONCLUSION

In this paper, we have proposed a network reconstructs the US volume using image features and optical flow features. To effectively use optical flow features, we introduced the pyramid warping layer which combines the optical flow features with image features using warping operation in different scales. We also adopted the fusion module with attention mechanism which fuses the warped features in different pyramid levels. What is more, we appended the channel and spatial attention which weights the features in channel and spatial dimension. The results of experiments validated that our reconstruction network had efficient performance on volume reconstruction which outperformed state-of-the-arts. It validated that the proposed network fusing optical flow features with warping operation in different scales performed better for the task of US volume reconstruction.

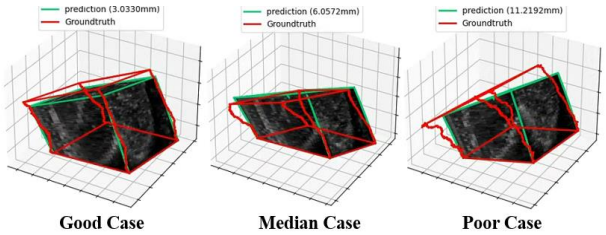

Fig. 3. 3D trajectory visualization of cases in testing data.

REFERENCES

[1] Guo, Hengtao, et al. "Transducer Adaptive Ultrasound Volume Reconstruction." arXiv preprint arXiv:2011.08419 (2020).

[2] Hennersperger, Christoph, et al. "Towards MRI-based autonomous robotic US acquisitions: a first feasibility study." IEEE transactions on medical imaging 36.2 (2016): 538-548.

[3] Busam, Benjamin, et al. "Markerless inside-out tracking for 3d ultrasound compounding." Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation. Springer, Cham, 2018. 56-64.

[4] Gee, Andrew H., et al. "Sensorless freehand 3D ultrasound in real tissue: speckle decorrelation without fully developed speckle." Medical image analysis 10.2 (2006): 137-149.

[5] Guo, Hengtao, et al. "Sensorless freehand 3D ultrasound reconstruction via deep contextual learning." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2020.

[6] Miura, Kanta, et al. "Localizing 2D Ultrasound Probe from Ultrasound Image Sequences Using Deep Learning for Volume Reconstruction." Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, Cham, 2020. 97-105.

[7] Prevost, Raphael, et al. "3D freehand ultrasound without external tracking using deep learning." Medical image analysis 48 (2018): 187-202.

[8] Tehrani, Ali KZ, and Hassan Rivaz. "Displacement estimation in ultrasound elastography using pyramidal convolutional neural network." IEEE transactions on ultrasonics, ferroelectrics, and frequency control 67.12 (2020): 2629-2639.

[9] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[10] Ilg, Eddy, et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[11] Dosovitskiy, Alexey, et al. "Flownet: Learning optical flow with convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

[12] Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[13] Feng, Shuanglang, et al. "CPFNet: Context pyramid fusion network for medical image segmentation." IEEE transactions on medical imaging 39.10 (2020): 3008-3018.

[14] Lasso, Andras, et al. "PLUS: open-source toolkit for ultrasound-guided intervention systems." IEEE transactions on biomedical engineering 61.10 (2014): 2527-2537.