Supplementary Material

Yikang Liu[1][0000−0003−1069−1215], Lin Zhao[1], Eric Z. Chen[1], Xiao Chen[1], Terrence Chen[1], and Shanhui Sun[1]

United Imaging Intelligence, Boston, MA, USA
shanhui.sun@uii-ai.com

1 Network Architectures

1.1 CNN-ConvLSTM

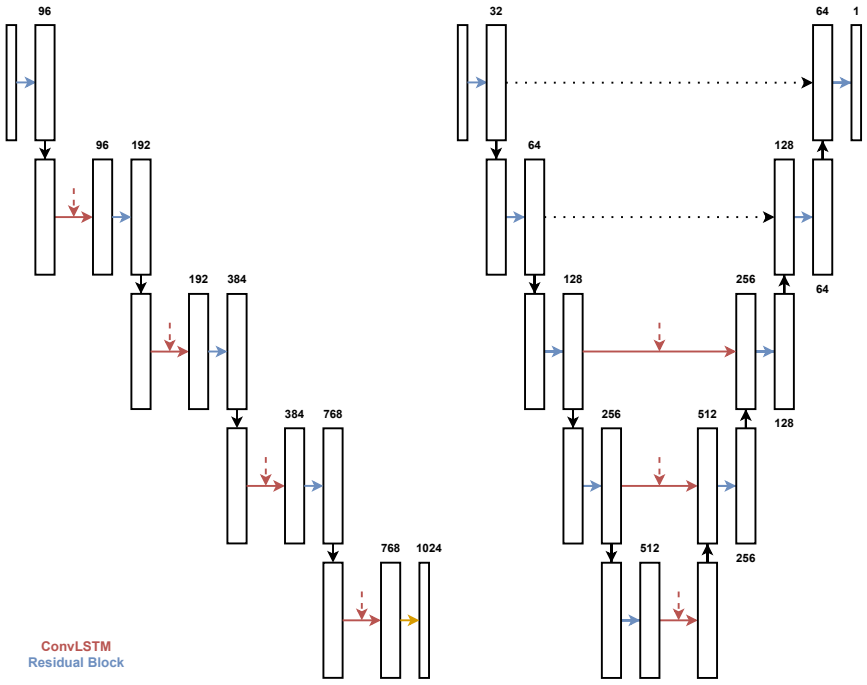

Fig. S1. Architectures of CNN-ConvLSTM Networks. The left is the model for cardiac phase matching, and the right is for catheter tip tracking. Numbers above tensors are channel numbers.

Fig. S1 shows the model architectures of CNN-ConvLSTM networks for cardiac phase matching (CPM) and catheter tip tracking (CTT).

The CPM model comprises a series of alternating UNetResBlock [1][1] layers, ConvLSTM [3][2] layers, and downsampling layers (implemented by 1x1 convolution layers with stride=2), followed by a global max pooling and a fully connected layer to transform a 4D image tensor into a 1D feature vector. A sequence of recorded cardiac angiographic images and a live fluoroscopic image stream are concatenated along the temporal dimension and sequentially fed into the model.

The CTT model is a UNet with ConvLSTM layers in the skip connections. The input is a sequence of 3-channel tensors, with each channel containing the reference image (the image where tip location is known), the reference tip heatmap, and the current image to inference. The tensors are sequentially inputted into the network, which then outputs a tip heatmap for each frame.

The input and output channel numbers of UNetResBlocks and ConvLSTM layers are shown in Fig. S1. Other hyperparameters are shown in Table S1.

Table S1. Hyperparameters of UNetResBlock and ConvLSTM Layers.

UNetResBlock		ConvLSTM	
spatial_dims	2	hidden_dim	output channel#
kernel_size	3	kernel_size	3
norm_name	None	bias	True
act_name	relu	layers	3 (the last) 1 (others)

1.2 CNN-Transformer

The CNN-Transformer model for cardiac phase matching comprises a ResNet encoder [2] (the part of ResNet-50 before the average pooling layer) and stacked attention layers. The outputs from the last two stages of the ResNet encoder are averaged globally and flattened and concatenated into 1D vectors (with a dimension of 3072) before being passed to the attention layers. Attention layers were implemented with *torch.nn.MultiheadAttention*[3], with *embed_dim*=3072, *num_heads*=4, and other parameters were set to defaults. Five attention layers were used with residual connections. The attention layers run with self-attention for recorded cardiac angiographic images. For real-time inference of the live fluoroscopic image stream, the features extracted from the current fluoroscopic image are used as the query vector, while features from previous frames are used as key and value vectors.

The CNN-Transformer model for cardiac tip tracking has the same backbone as STARK-S50[4], except for that one heatmap was generated indicating the tip location instead of two heatmaps for corners of the bounding box. The template size is 64x64.

[1] https://docs.monai.io/en/stable/networks.html, v1.2.0
[2] https://github.com/ndrplz/ConvLSTM_pytorch
[3] v2.0

References

1. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
4. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10448–10457 (2021)