HHS Public Access

Author manuscript

Med Image Comput Comput Assist Interv. Author manuscript; available in PMC 2024 October 28.

Published in final edited form as:

Med Image Comput Comput Assist Interv. 2024 October ; 15007: 701–711.

doi:10.1007/978-3-031-72104-5_67.

Tagged-to-Cine MRI Sequence Synthesis via Light Spatial-Temporal Transformer

Xiaofeng Liu[1,2], **Fangxu Xing**[1], **Zhangxing Bian**[3], **Tomas Arias-Vergara**[1,4], **Paula Andrea Pérez-Toro**[1,4], **Andreas Maier**[4], **Maureen Stone**[5], **Jiachen Zhuo**[5], **Jerry L. Prince**[3], **Jonghye Woo**[1]

[1]Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[2]Yale University, New Haven, CT, USA

[3]Johns Hopkins University, Baltimore, MD, USA

[4]Friedrich-Alexander University, Erlangen, Germany

[5]University of Maryland, Baltimore, MD, USA

Abstract

Tagged magnetic resonance imaging (MRI) has been successfully used to track the motion of internal tissue points within moving organs. Typically, to analyze motion using tagged MRI, cine MRI data in the same coordinate system are acquired, incurring additional time and costs. Consequently, tagged-to-cine MR synthesis holds the potential to reduce the extra acquisition time and costs associated with cine MRI, without disrupting downstream motion analysis tasks. Previous approaches have processed each frame independently, thereby overlooking the fact that complementary information from occluded regions of the tag patterns could be present in neighboring frames exhibiting motion. Furthermore, the inconsistent visual appearance, e.g., tag fading, across frames can reduce synthesis performance. To address this, we propose an efficient framework for tagged-to-cine MR sequence synthesis, leveraging both spatial and temporal information with relatively limited data. Specifically, we follow a split-and-integral protocol to balance spatialtemporal modeling efficiency and consistency. The light spatial-temporal transformer (LiST^2) is designed to exploit the local and global attention in motion sequence with relatively lightweight training parameters. The directional product relative position-time bias is adapted to make the model aware of the spatial-temporal correlation, while the shifted window is used for motion alignment. Then, a recurrent sliding fine-tuning (ReST) scheme is applied to further enhance the temporal consistency. Our framework is evaluated on paired tagged and cine MRI sequences, demonstrating superior performance over comparison methods.

1 Introduction

In muscle mechanic-related medical imaging applications, assessment of deformation patterns of internal tissues is a crucial step that translates image-related information to

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

physical spaces. The need for more accurate estimations is essential in both imaging research and clinical practice. Tagged magnetic resonance imaging (MRI) [18] with spatially encoded tag patterns is among the most popular modalities for capturing internal motion in terms of deformed image patterns, which has received widespread applications in studying the motion of internal muscles, such as the myocardium [16] and the tongue [24,23].

However, due to the intrinsic low anatomical resolution of tagged MRI, an additional set of cine MRI scans with higher resolution are often acquired to serve as a matching pair that specifies anatomical structures in detail [17]. They are often used in localization, segmentation, and constraining of the tagged images [7], albeit at the cost of extra time and expense. Recent studies propose various methods for tagged-to-cine MR translation [11,13,12,9], aiming to generate cine MRI sequences from tagged MRI data. By utilizing these generated cine MRI images, the need for acquiring additional cine MRI scans is obviated, thereby almost reducing acquisition time by a half, while maintaining the workflow for subsequent motion estimation analyses. However, the existing methods [11,13,12,9] fail to fully exploit the video nature of collected data since they process frames independently, disregarding the complementary information embedded in neighboring frames, which has been well demonstrated in nature video restoration [20,5,8]. In addition, the temporal flickering with inconsistent visual appearance in neighboring frames can distract the viewers and result in incoherent results for the subsequent motion analysis. Notably, the fading of tags across input frames can further reduce synthesis coherence.

A key aspect of video restoration methods lies in designing components to realize alignment across frames for exploiting inter-frame complementary information. Some sliding window-based methods [20,15] only use CNNs without explicit alignment. These methods generally input a short video section to extract temporal information implicitly, but their performance tends to be suboptimal. Optical flow [19] or deformable convolution [26] is usually required, at the cost of a more complex computational architecture and increased time expense. Some other methods are based on a recurrent algorithm [1]; however, they suffer from significant performance drops when processing either short or long videos [8]. Recently, vision transformer (ViT) has become a promising alternative for attaining long-range receptive fields [8]. However, its long-input setting can significantly reduce the available training inputs, and its costly self-attention mechanism cannot be adequately supported by the limited data.

To address the aforementioned challenges, in this work, we propose a splitand-integral protocol to balance spatial-temporal modeling efficiency and consistency. Specifically, we propose a **li**ght **s**patial-**t**emporal **t**ransformer (LiST^2) that adapts the image-based LightViT [4] to the video sequence for efficient local and global spatial-temporal modeling. We equip it with a directional product relative position-time bias to make the model aware of the spatial-temporal correlation. In addition, the shifted window for local patch design is also applied to address the motion-related mismatch in the boundary regions of neighboring frames. Therefore, our LiST^2 inherits the efficiency of sliding window-based methods, while being able to explicitly explore spatial-temporal correlations. However, ensuring consistency at intersections remains uncertain. To address this, we propose a **re**current **s**liding fine-**t**uning (ReST) scheme, which is regularized by the overlap-consistent loss. With this add-on

step, we are able to combine the sliding window model with a recurrent algorithm, which strengthens the interaction between temporal sections without compromising spatial quality.

The main contributions of this work are three-fold:

- To our knowledge, this is the first attempt at tagged-to-cine MRI sequence synthesis, which explicitly explores the complementary cross-frame information.
- We developed a light spatial-temporal transformer with position-time bias and shifted window to achieve efficient motion modeling with relatively limited data.
- We further explored a recurrent sliding scheme as a fine-tuning phase to strengthen temporal consistency between neighboring sliding sections.

Both quantitative and qualitative evaluation results from 20 healthy controls with a total of 3,774 paired slices of tagged and cine MRI show the validity of our proposed LiST^2 + ReST framework and its superiority to conventional image and long-video-based translation methods.

2 Methodology

Given the paired tagged and cine MRI sequences, we adopt a split-and-integral protocol with LiST^2 and ReST to balance data efficiency and temporal consistency. Following the temporal sliding windows [20,21], we segment a long video sequence $\mathbf{X}$ into several overlapping sections $\mathbf{x}^s \in \mathbb{R}^{H \times W \times T}$, each with a fixed length T, where s indexes the section. We empirically set T to consist of five consecutive frames $\left\{\mathbf{x}_t^s \in \mathbb{R}^{H \times W \times 1}\right\}_{t=1}^{T}$ [20], and we use a step size of $T//2$ frames[6] to address the correlation at intersections in our ReST phase.

Following the conventional approach used in ViT for video restoration and processing [8,10], each frame $\mathbf{x}_t^s$ is processed by two layers of shared 2D convolution, serving as the encoder. This facilitates efficient modeling of neighboring correlations to extract features $\mathbf{f}_t^s \in \mathbb{R}^{C_l \times H_l \times W_l \times 1}$ after the l-th layer, followed by applying an attention scheme. Importantly, we employ a shared decoder architecture based on UNet, incorporating deconvolution layers and skip connections. This decoder is used to synthesize the corresponding cine MRI sequences $\left\{\tilde{\mathbf{c}}_t^s \in \mathbb{R}^{H \times W \times 1}\right\}_{t=1}^{T}$, approximating the ground truth $\{\tilde{\mathbf{c}}_t^s\}_{t=1}^{T}$.

2.1 Light Spatial-Temporal Transformer (LiST^2)

Directly modeling spatial-temporal correlations among any two C_l-dim vectors in $H \times W$ plane of $\mathbf{f}_t^s$ can impose quadratic complexity of $\mathcal{O}\left(H_l^2 W_l^2 T^2\right)$. The recent efficient ViTs [4,25,2] usually adopt a local patch design to compute local selfattention and correlate global patches with CNNs or anchored global attention. We further extend the general idea to the 3D volume sequence. Specifically, in the $H \times W$ plane, the feature $\mathbf{f}_t^s$ is divided

[6] //: floor division, which rounds the division result down to the nearest integer.

into non-overlap local patches with the size of $M_H \times M_W$. For a spatial-temporal patch, we concatenate T frames to have $N = [(M_H \times M_W) \times T]$ tokens, which are processed by the linear projections of query (W_q), key (W_k), and value (W_v) branches with d-dim output to have a 2D matrix $\mathbf{f}_i^q, \mathbf{f}_i^k, \mathbf{f}_i^v \in \mathbb{R}^{N \times d}$ [3,25,2]. The i-th local patch self-attention across T frames is then formulated as

$$\mathbf{f}_i^{\text{local}} = \text{Attn}\left(\mathbf{f}_i^q, \mathbf{f}_i^k, \mathbf{f}_i^v\right) = \text{SoftMax}\left(\frac{\mathbf{f}_i^q \mathbf{f}_i^{k\top} + \mathbf{r}_i}{\sqrt{d}}\right)\mathbf{f}_i^v, \in \mathbb{R}^{N \times d}. \tag{1}$$

To enable the model to be aware of spatial-temporal correlations, we propose adding a directional product relative position-time bias $\mathbf{r}_i \in \mathbb{R}^{N \times N}$, where element $r_{a,b} = \mathbf{p}_{\delta_{a,b}^H, \delta_{a,b}^W, \delta_{a,b}^T}, \in \mathbb{R}$ is the relative position weight between the voxel a and b in $H_l \times W_l \times T$ tensor[7]. Notably, while the bias as in [22] only addresses spatial considerations, temporal aspects are crucial for our task. We set the offset of patch positions in three directions as follows: $\delta_{a,b}^H = H_a - H_b + H_l$, $\delta_{a,b}^W = W_a - W_b + W_l$, and $\delta_{a,b}^T = T_a - T_b + T$ as the indices in the learnable tensor $\mathbf{p}$. Therefore, our $\mathbf{r}_i$ can distinguish between spatial or temporal offsets and their direction.

In addition, motion can introduce offsets between corresponding pixels in different sequential frames, and boundary pixels may often transition across pre-defined patches spanning multiple frames. Therefore, using a fixed patch splitting or pyramid shrinking of the patch [4,25,2] may not efficiently explore cross-frame information in boundary regions. As such, we propose to adopt the approach in SWinTransformer [14] to different splitting methods within consecutive transformer blocks to provide connections within various possible patch combinations. In addition, as in [14], we set the regular patch window with fixed dimensions of $M_H = M_W = 4$, and displace the shifted window by (2,2) pixels. By doing this, no pixel will always be the middle patch boundary. Thus, with a constant N, the computational complexity of local processing for all patches becomes $\mathcal{O}\left(N^2 \frac{H_l}{M_H} \frac{W_l}{M_W}\right)$, which scales linearly with the input spatial size.

For efficient global correlation, we follow the lightweight transformer design [4], which uses a global embedding $\mathbf{G} \in \mathbb{R}^{K \times d}$ with $K \ll H_l \times W_l \times T$ randomly generated global tokens as the anchor for global information aggregation $\hat{\mathbf{G}}$. We perform the aggregation using the attention mechanism with $\mathbf{G}^{\mathbf{q}}, \mathbf{f}_\mathbf{i}^{\mathbf{k}}$, and $\mathbf{f}_\mathbf{i}^{\mathbf{v}}$, which is then broadcasted with the attention mechanism of $\mathbf{f}_\mathbf{i}^{\mathbf{q}}, \hat{\mathbf{G}}^{\mathbf{k}}$, and $\hat{\mathbf{G}}^{\mathbf{v}}$ to leverage global contextual information [4]. Notably, with a global reference anchor $\mathbf{G}$, we can process each patch in parallel. Thus, for global attention, information from local tokens can be aggregated by modeling their global dependencies using

[7] A learnable tensor $\mathbf{p} \in \mathbb{R}^{(2H_l - 1) \times (2W_l - 1) \times (2T - 1)}$ is initialized with trunc_normal_ [22], where H_l and W_l are the maximum patch dimensions in **global** attention. Of note, we use the same $\mathbf{p}$ for local and global attention.

$$\hat{\mathbf{G}}_i = \mathrm{Attn}\left(\mathbf{G}^q, \mathbf{f}_i^k, \mathbf{f}_i^v\right) = \mathrm{SoftMax}\left(\frac{\mathbf{G}^q \mathbf{f}_\mathbf{i}^{\mathbf{k}\top} + \mathbf{r}_i}{\sqrt{d}}\right)\mathbf{f}_i^v, \in \mathbb{R}^{N \times d}. \tag{2}$$

We then broadcast these global dependencies to each local token as follows:

$$\mathbf{f}_i^{\mathrm{global}} = \mathrm{Attn}\left(\mathbf{f}_i^q, \widehat{\mathbf{G}}_i^k, \widehat{\mathbf{G}}_i^v\right) = \mathrm{SoftMax}\left(\frac{\mathbf{f}_i^q \widehat{\mathbf{G}}_\mathbf{i}^{\mathbf{k}\top} + \mathbf{r}_i}{\sqrt{d}}\right)\widehat{\mathbf{G}}_i^v, \in \mathbb{R}^{N \times d}. \tag{3}$$

For each patch, we have the final feature $\mathbf{f}_i' = \mathbf{f}_i^{\mathrm{local}} + \mathbf{f}_i^{\mathrm{global}}, \in \mathbb{R}^{N \times d}$. By adding $\mathbf{f}_i^{\mathrm{local}}$ and $\mathbf{f}_i^{\mathrm{global}}$, each token can leverage both local and global features while maintaining linear complexity relative to the input size. This results in noticeable improvements with a negligible increase in FLOPs.

After the processing of our L blocks of LiST^2, deconvolution with skip connections is applied to generate the cine MRI sequences. The model is trained with the following reconstruction loss:

$$\mathscr{L}_{\mathrm{rec}}^{LiST^2} = \| \left\{\tilde{c}_{-2}^{s+1}, \tilde{c}_{-1}^{s+1}, \tilde{c}_0^{s+1}, \tilde{c}_1^{s+1}, \tilde{c}_2^{s+1}\right\}, \left\{\left\{\tilde{c}_{-2}^{s+1}, \tilde{c}_{-1}^{s+1}, \tilde{c}_0^{s+1}, \tilde{c}_1^{s+1}, \tilde{c}_2^{s+1}\right\}\right\} \|_2^2 . \tag{4}$$

Note that with $T = 5$ and a step size of $T//2 = 2$, we use only three middle generated frames from each video section as our final results.

2.2 Recurrent Sliding Fine-Tuning (ReST)

To improve coherence between temporal sections, we further adapt the recurrent scheme as a fine-tuning phase with additional loss regularization to alleviate flickering. Specifically, we input the frame of section $s + 1$ as the last output frame of section s, which can be described as follows:

$$\text{Section } s+1: \left\{\tilde{c}_{-1}^{s+1}, \tilde{c}_0^{s+1}, \tilde{c}_1^{s+1}\right\} = \mathrm{LiST}^2\left(\left\{\tilde{c}_0^s, x_{-1}^{s+1}, x_0^{s+1}, x_1^{s+1}, x_2^{s+1}\right\}\right). \tag{5}$$

We calculate $\mathscr{L}_{rec}^{ReST} = \| \left\{\tilde{c}_{-1}^{s+1}, \tilde{c}_0^{s+1}, \tilde{c}_1^{s+1}\right\}, \left\{\tilde{c}_{-1}^{s+1}, \tilde{c}_0^{s+1}, \tilde{c}_1^{s+1}\right\} \|_2^2$ as our reconstruction loss. Moreover, an overlap exists in the output frames of the neighboring sections. Therefore, we enforce the consistency with $\mathscr{L}_{overlap} = \| \tilde{c}_1^s - \tilde{c}_{-1}^{s+1} \|_2^2$. The overall loss in ReST phase is $\mathscr{L} = \mathscr{L}_{rec} + \alpha \mathscr{L}_{overlap}$.

3 Experiments and Results

For the experiments carried out in this work, 20 sequences with a total of 3,774 paired tagged and cine MRI frames were acquired from a total of 20 healthy controls, while speaking an utterance, “a souk” [7,23]. The image sequence was acquired at a rate of 26 frames per second. Both cine and tagged MRI images are in the same spatiotemporal coordinate space. For our task, tagged MRI images with horizontal tag patterns were utilized. Each frame of tagged and cine MRI has a spatial size of 128×128.

We employed a subject-independent five-fold cross-validation approach. In each fold, one subject was used for testing, while the remaining four subjects were used for training and validation. The long video was divided into sections consisting of five frames with a step size of two. For data augmentation, we applied random jitter by resizing the input images from 128 × 128 to 140 × 140, followed by random cropping back to the original size of 128 × 128 similar to [9,6].

We utilized the standard UNet backbone, employing two convolutional or deconvolutional blocks in our encoder and decoder modules, respectively. We note that all compared methods consistently utilized the L2 minimization objective. Empirically, we set $L = 4, K = 64$, and $\alpha = 1$. Our framework was implemented using the PyTorch deep learning toolbox, with a learning rate set to 1e-4 in the Adam optimizer. The training was conducted on an NVIDIA A100 GPU, requiring approximately 5 hours for 200 epochs of LiST^2 training, followed by 100 epochs of ReST. During testing, translating one tagged MRI section to the corresponding cine MR images took about 0.2 seconds.

We anticipate that the synthesized images will demonstrate realistic and structurally consistent textures relative to their corresponding ground truth images, which is essential for subsequent analyses. In Fig. 2, we present a qualitative comparison between our proposed LiST^2 and ReST methods. We can clearly observe that the generated cine MRI sequence aligns well with the ground truth, achieving superior structural and texture consistency among neighboring frames. In the LiST^2-only model, i.e., without the ReST phase, we can achieve relatively good temporal consistency within each section. However, flickering may still be present in the outputs across different sections. Additionally, the reconstructed tongue shape in the second section also shows a relatively large deformation compared with the ground truth. For LiST^2 w/o shifted window (SW), the generated results can have a large degradation. SW can be an important module in our framework to enhance the local attention of moving parts. Without SW, the spatial-temporal correlation might be disrupted by mismatches in pixels caused by motion. The considerable variation in contrast and texture within the tongue region across real cine MRI images could potentially impact subsequent analyses. Furthermore, without a lightweight design, directly applying the SWinViT [14] to video data may not yield visually satisfactory results. The relatively limited dataset may not adequately support the training of such a large model. While the framewise method [11] is able to synthesize cine MRI sequences with good visual quality, coherence among neighboring frames may be lacking. For example, the second frame in Fig. 2 often appears considerably smoother compared with the third frame, leading to distortion

in the anatomical structure. Notably, the use of adversarial loss as in [11] does not lead to improved performance in our video synthesis task.

To quantitatively evaluate our framework, we employed established evaluation metrics, such as the Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean L1 error [11,9]. Table 1 lists numerical comparisons between the proposed framework, image-based tagged-to-cine methods (e.g., UNet [9] and Bi-VAE-GAN [11]), the recent CNN-based sliding window method DVDnet [20] with five-frame section, as well as video-based transformer VRT [8]. The proposed LiST^2 outperformed the other three comparison methods with respect to L1 error, SSIM, and PSNR consistently. By integrating five consecutive frames as input, DVDnet [20] with CNN-based sliding window can achieve improvements over the frame-based methods [9,11], while its spatial-temporal modeling ability could be inferior to the ViT-based models [8]. However, the relatively limited video data cannot adequately support the VRT with a 16-frame video section, leading to inferior performance. Similarly, the absence of a lightweight ViT design, such as replacing the LiST^2 module with SWinViT [14], also results in a performance drop. Furthermore, the performance can be enhanced even further through post-training fine-tuning with ReST. It is worth noticing that recurrent-only models often experience significant performance degradation when applied to short sequences [8], because of the lack of enough long sequences for training. As shown in Fig. 3, the relatively high performance of LiST^2 can be achieved with $L \in [4,\ 6\ ]$, $T \in [5,\ 7\ ]$ and $K \in [64, 128]$. The lower one is usually adopted for efficiency. In addition, the weight of $\mathscr{L}_{overlap}$ could be $\alpha \in [0.5, 2]$.

4 Conclusions

In this work, we proposed to synthesize cine MRI sequences from its paired tagged MRI sequences. Given that occluded information by the tag patterns can be inherited in neighboring frames, it is advantageous to leverage inter-frame information. In addition, temporal flickering is a long-lasting challenge in sequence processing. With relatively limited data, we proposed a split-and-integral protocol with LiST^2 and ReST to balance spatial-temporal modeling efficiency and consistency. We systematically designed a lightweight LiST^2 framework to achieve video translation, in which the directional product relative position-time bias and shifted window were further adapted to catering the motion. The ReST phase can be a general module to be added on sliding window-based methods for improved cross-section consistency. Our experimental results showed that our method outperformed the comparison methods both quantitatively and qualitatively. The synthesized cine MRI holds promise for further applications, such as tongue segmentation and surface motion observation.

Acknowledgements

This work is supported by NIH R01DC018511, R01DC014717, R01CA133015, and R21EB034911.

References

1. Chan KC, Wang X, Yu K, Dong C, Loy CC: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4947–4956 (2021)
2. Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, Xia H, Shen C: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems 34, 9355–9366 (2021)
3. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. : An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Huang T, Huang L, You S, Wang F, Qian C, Xu C: Lightvit: Towards light-weight convolution-free vision transformers. arXiv preprint arXiv:2207.05557 (2022)
5. Isobe T, Jia X, Gu S, Li S, Wang S, Tian Q: Video super-resolution with recurrent structure-detail network. In: Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 645–660. Springer (2020)
6. Isola P, Zhu JY, Zhou T, Efros AA: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
7. Lee J, Woo J, Xing F, Murano EZ, Stone M, Prince JL: Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. In: ISBI. pp. 1465–1468. IEEE (2013)
8. Liang J, Cao J, Fan Y, Zhang K, Ranjan R, Li Y, Timofte R, Van Gool L: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)
9. Liu X, Prince JL, Xing F, Zhuo J, Reese T, Stone M, El Fakhri G, Woo J: Attentive continuous generative self-training for unsupervised domain adaptive medical image translation. Medical Image Analysis p. 102851 (2023) [PubMed: 37329854]
10. Liu X, Xing F, Prince J, Stone M, El Fakhri G, Woo J: Synthesizing audio from tongue motion during speech using tagged mri via transformer. In: Medical Imaging 2023: Image Processing. vol. 12464, pp. 203–207. SPIE (2023)
11. Liu X, Xing F, Prince JL, Carass A, Stone M, El Fakhri G, Woo J: Dualcycle constrained bijective vae-gan for tagged-to-cine magnetic resonance image synthesis. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1448–1452. IEEE (2021)
12. Liu X, Xing F, Prince JL, Stone M, El Fakhri G, Woo J: Structure-aware unsupervised tagged-to-cine mri synthesis with self disentanglement. In: Medical Imaging 2022: Image Processing. vol. 12032, pp. 470–476. SPIE (2022)
13. Liu X, Xing F, Stone M, Zhuo J, Reese T, Prince JL, El Fakhri G, Woo J: Generative self-training for cross-domain unsupervised tagged-to-cine mri synthesis. In: International Conference on Medical Image Computing and ComputerAssisted Intervention. pp. 138–148. Springer (2021)
14. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Maggioni M, Huang Y, Li C, Xiao S, Fu Z, Song F: Efficient multistage video denoising with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3466–3475 (2021)
16. Osman NF, McVeigh ER, Prince JL: Imaging heart motion using harmonic phase mri. TMI 19(3), 186–202 (2000)
17. Parthasarathy V, Prince JL, Stone M, Murano EZ, NessAiver M: Measuring tongue motion from tagged cine-mri using harmonic phase (harp) processing. The Journal of the Acoustical Society of America 121(1) (2007)
18. Petitjean C, Rougon N, Cluzel P: Assessment of myocardial function: a review of quantification methods and results using tagged mri. Journal of Cardiovascular Magnetic Resonance (2005)
19. Shi X, Huang Z, Bian W, Li D, Zhang M, Cheung KC, See S, Qin H, Dai J, Li H: Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. arXiv preprint arXiv:2303.08340 (2023)

20. Tassano M, Delon J, Veit T: Fastdvdnet: Towards real-time deep video denoising without flow estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1354–1363 (2020)
21. Wang C, Zhou SK, Cheng Z: First image then video: A two-stage network for spatiotemporal video denoising. arXiv preprint arXiv:2001.00346 (2020)
22. Wu K, Peng H, Chen M, Fu J, Chao H: Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10033–10041 (2021)
23. Xing F, Woo J, Lee J, Murano EZ, Stone M, Prince JL: Analysis of 3-D tongue motion from tagged and cine magnetic resonance images. Journal of Speech, Language, and Hearing Research 59(3), 468–479 (2016)
24. Xing F, Woo J, Gomez AD, Pham DL, Bayly PV, Stone M, Prince JL: Phase vector incompressible registration algorithm for motion estimation from tagged magnetic resonance images. IEEE TMI 36(10) (2017)
25. Zhang Q, Yang YB: Rest: An efficient transformer for visual recognition. Advances in Neural Information Processing Systems 34, 15475–15485 (2021)
26. Zhu X, Hu H, Lin S, Dai J: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)

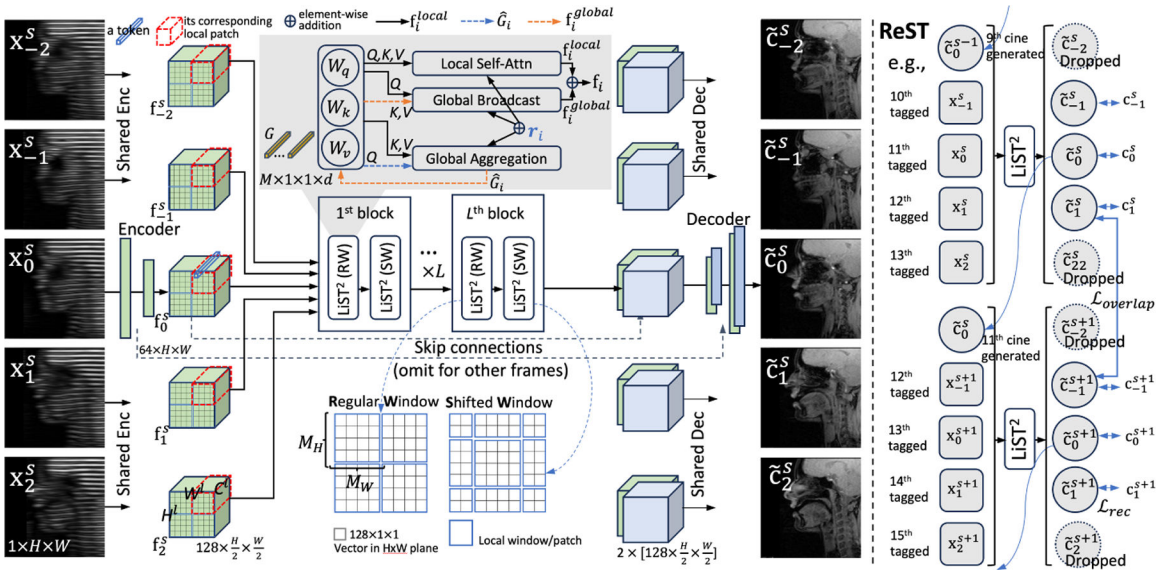

Fig. 1:

Illustration of our video tagged-to-cine synthesis framework. Left: the UNet model with LiST^2 module. Right: ReST phase tuning with overlap loss.

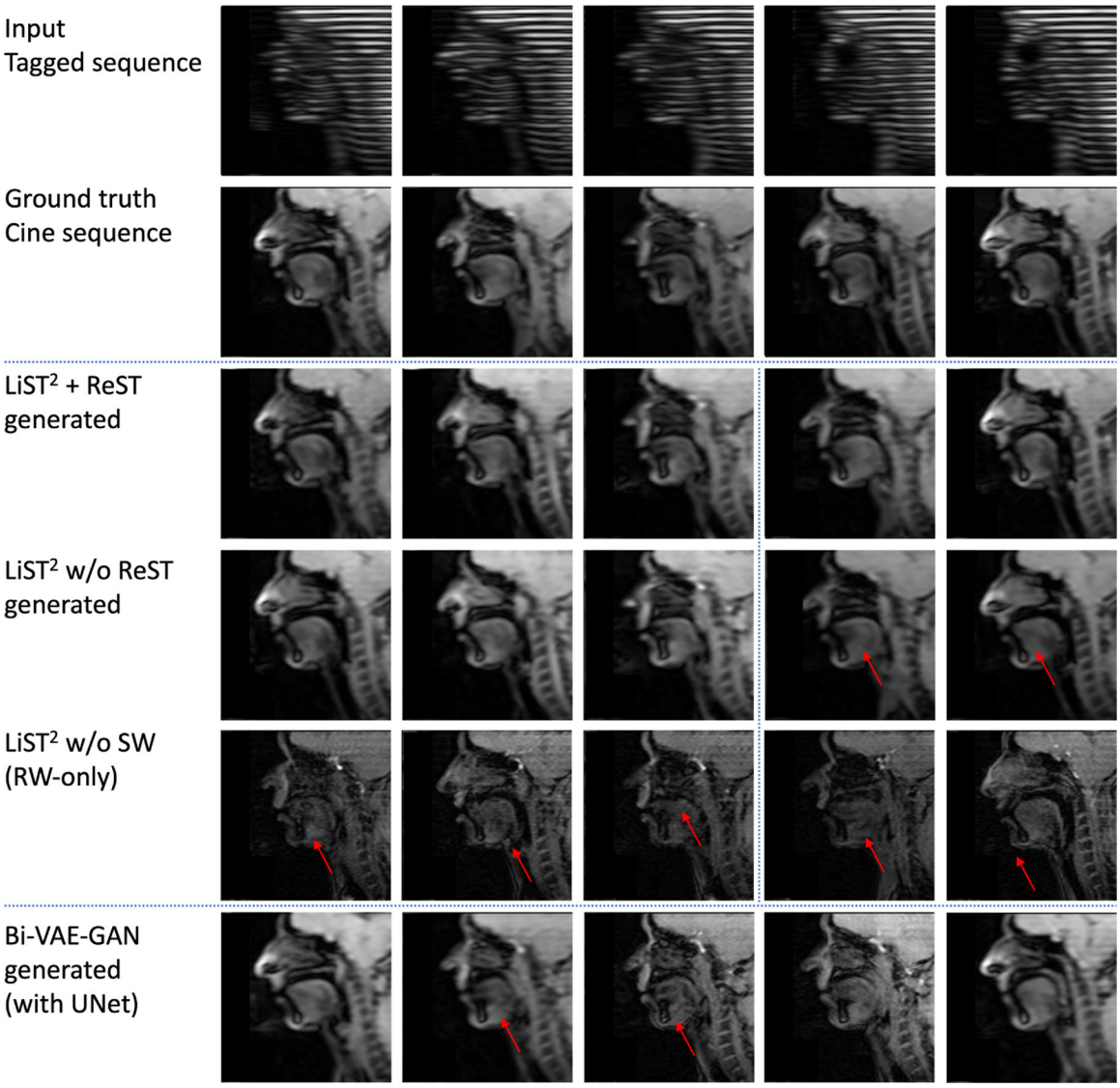

Fig. 2:

Qualitative comparisons of our proposed LiST^2 + ReST with imagebased Bi-VAE-GAN (UNet) [11] and ablation study. The first three and later two frames are from two consecutive sections in LiST^2-based methods.

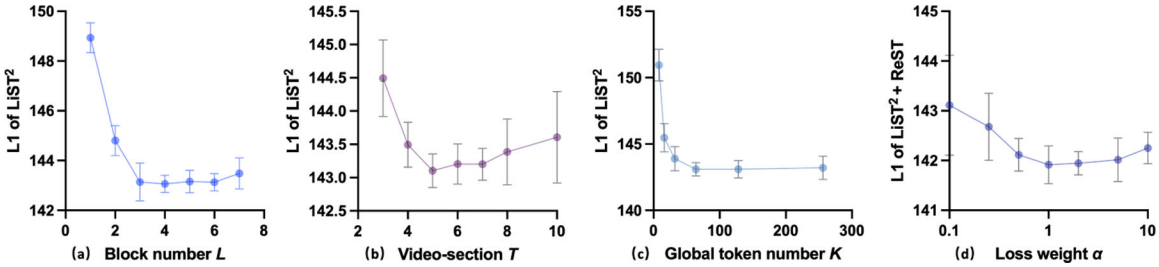

Fig. 3:

Sensitivity analysis of our LiST^2: (a) block number L, (b) section length T, and (c) global token number K, as well as (d) $\mathscr{L}_{\text{overlap}}$ weight α used in ReST.

Table 1:

Numerical comparisons and ablation study with five-fold crossevaluation. The best results are shown in **bold**. We report the results as mean±SD over three random initializations.

Methods	**Processing input**	**L1 ↓**	**SSIM ↑**	**PSNR ↑**
UNet [9]	Image	157.53±0.14	0.9426±0.0034	32.85±0.12
Bi-VAE-GAN (UNet) [11]	Image	152.47±0.28	0.9502±0.0074	36.48±0.30
DVDnet [20]	Sliding section	150.24±0.15	0.9678±0.0031	36.46±0.15
VRT [8]	Video (16-frame)	208.71±0.18	0.8742±0.0023	16.97±0.19
Proposed (LiST^2 + ReST)	Sliding section	**141.91**±0.18	**0.9743**±0.0035	**38.95**±0.15
Proposed (LiST^2 only)	Sliding section	143.05±0.20	0.9728±0.0027	38.72±0.18
Proposed (LiST^2 w/o SW)	Sliding section	147.68±0.26	0.9715±0.0025	38.49±0.16
Proposed (Replace LiST^2 to SWinViT [14])	Sliding section	192.54±0.22	0.9135±0.0042	30.54±0.22