CardioInterp: Generative Modeling for Cardiovascular OCT Interpolation with Anatomical Continuity and Fidelity

Linyuan Li[1[0009−0006−5833−526X]], Bing Yang[1[0000−0003−1234−2916]], Minqing Zhang[1[0000−0002−7214−0569]], Mengxian He[1[0009−0005−7585−374X]], and Wu Yuan[1✉[0000−0001−9405−519X]]

The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong
wyuan@cuhk.edu.hk

Abstract. Cardiovascular Optical Coherence Tomography (OCT) is hindered by the brief imaging window provided by contrast agents, making it challenging to capture high-resolution images of multiple plaques over long vessel sections. Rapid catheter pullback and coarse spatial resolution increase the likelihood of missing subtle pathologies and critical plaque microstructures, compromising diagnostic accuracy. To address this, we introduce *CardioInterp*, the first generative interpolation model for cardiovascular OCT, designed to synthesize high-fidelity intermediate B-slices, enhancing structural continuity and spatial resolution. Our architecture integrates a latent diffusion framework with a novel Dual-Path Fusion Decoder designed to ensure inter-slice structural continuity while preserving microanatomical fidelity. Experiments on cardiovascular OCT datasets demonstrate that *CardioInterp* achieves superior interpolation quality (PSNR=28.59, SSIM=51.80%) at 6 times upscaling of B-slices and spatial resolution, surpassing traditional medical image interpolation methods and setting a new benchmark. This innovative computational approach enables high-resolution imaging of long vessel sections within a limited temporal window in cardiovascular OCT. The code is available at: https://github.com/Lee728243228/CardioInterp.

Keywords: Medical imaging · Cardiovascular OCT · Intermediate B-slice · Diffusion model · Medical slice interpolation.

1 Introduction

Optical Coherence Tomography (OCT), characterized by its micrometer-scale resolution and real-time volumetric imaging capability [15], has become an essential diagnostic and therapeutic intervention for cardiovascular diseases, whose data collection method is shown in Fig. 1. Nevertheless, two inherent limitations exist. The first is the rapid cardiac motion, ranging from 60 to 100 beats per minute. The second is the short effective dwell time of intravascular contrast agents, approximately 1 to 3 seconds. These factors compromise imaging completeness. To ensure full vascular coverage under these constraints, clinicians

must operate catheters at high pullback speeds during image acquisition, whose speeds generally range from 20 to 40 mm/s. [1], creating significant spatial discontinuities between adjacent B-slices, with spacing of approximately 100 to 200 µm, severely degrading the integrity of vascular wall microstructures in reconstructed 3D volumes and amplify diagnostic uncertainties [16]. Consequently, bridging the resolution-completeness trade-off through computational B-slice interpolation has become a pressing unmet need for achieving precision-guided cardiovascular interventions.

Given the specialized dynamic imaging paradigm characteristic of cardiovascular OCT system, cardiovascular OCT exhibits uniquely anisotropic resolution characteristics: B-scan spacing of $2 \times 10^{-1}mm$ versus $7 \times 10^{-3}mm$ between adjacent A-lines,which is different from common medical modalities such as CT and MRI's $0.5mm$ coronal/sagittal resolution and $1 \sim 3mm$ axial slices. Current medical image interpolation methodologies predominantly target CT/MRI protocols [2, 14, 9, 11], yet inadequately address cardiovascular OCT's marked anisotropy-posing technical hurdles in cross-slice continuity and intra-slice fidelity preservation during volumetric reconstruction.

Thus, the pivotal challenge for reconstructing intermediate OCT slices is establishing dynamically coherent representations under the anisotropic imaging constraints inherent to cardiovascular OCT systems while requiring structural continuity maintenance and feature fidelity preservation.

Large-motion interpolation [13, 5] exhibits similarities with highly anisotropic medical slice interpolation, sharing the challenge of significant inter-frame variations. The implementation of diffusion models in modeling large inter-frame motion demonstrates their robust capacity in capturing continuous feature transitions across substantial deformations—particularly evident in applications requiring preservation of structural coherence under large geometrical discrepancies. For example, VIDIM [5] generates highly realistic inter-frames, outperforming other traditional flow-based video frames interpolation models.

Based on the aforementioned analysis, we propose *CardioInterp*-the first generative interpolation model for cardiovascular OCT B-slice sequences, designed to synthesize high-fidelity intermediate slices while enhancing vascular structural continuity. Our method is developed based on latent scan interpolation diffusion model, that leveraging the 3D representation capacity and large-motion modeling capabilities of diffusion models, providing a robust generative foundation for the strong spatial anisotropy inherent to cardiovascular OCT. Further, we introduce an innovative dual-path fusion decoder that employs a dual-path fusion strategy to synergistically integrate deep semantic features and shallow texture features during the decoding phase, ensuring both fine-grained detail preservation and structural continuity in generated B-slices. Additionally, we incorporate Temporal Shift Module (TSM) [8] as the Pseudo-3D (P-3D) Block, a computational optimization strategy that enhances slice-to-slice continuity without introducing extra parameters or computational overhead. In summary, our contributions are:

1. We are the first focus on cardiovascular OCT interpolation, providing an innovative approach enabling high-resolution imaging of long vessel.

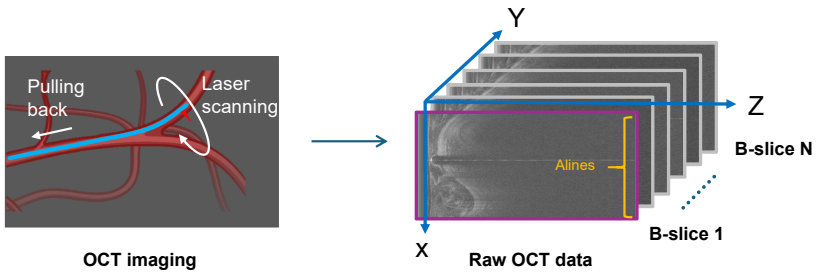

Fig. 1. OCT data collection and structure.

2. We propose *CardioInterp*, a latent scan interpolation diffusion model together with a novel dual-path fusion decoder, which utilize a dual path fusion strategy to compensate for the generation continuity and details.
3. Comprehensive experiments demonstrate that *CardioInterp* manages to generate intermediate slice with anatomical continuity and fidelity and achieve the state-of-the-art performance in both PSNR and SSIM.

2 Method

CardioInterp aims to generate intermediate scans with high fidelity and vascular structural continuity. As illustrated in Fig. 2, it operates under the DDPM [4, 12, 3] paradigm, integrating binary masks and latent representations of the first and final scans as conditional constraints to achieve interpolation via conditional video generation [7]. Furthermore, we utilize a dual-path fusion decoder that combines deep-level semantic and shallow-level texture features to enhance both spatial continuity and microstructural fidelity in reconstructed scans.

2.1 Latent Scan Interpolation Diffusion Model

2D Autoencoder for OCT Compression *CardioInterp* employs a lightweight 2D autoencoder tailored for interpolation tasks. This module compress OCT scan sample $\mathbf{x}_0 \sim p_{OCT}(\mathbf{x}_0)$ where $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times L\times 1}$ into the latent space to extract image latent features $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ where $\mathbf{z}_0 \in \mathbb{R}^{h\times w\times L\times c}$, $h = H/f$ and $w = W/f$. f indicates a spatial downsampling factor.

Conditional DDPM for Interpolation For training the conditional diffusion model, the gaussian noise is gradually added to the compressed latent code $\mathbf{z}_0$ through a diffusion process for t steps, which can be expressed as:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{1}$$

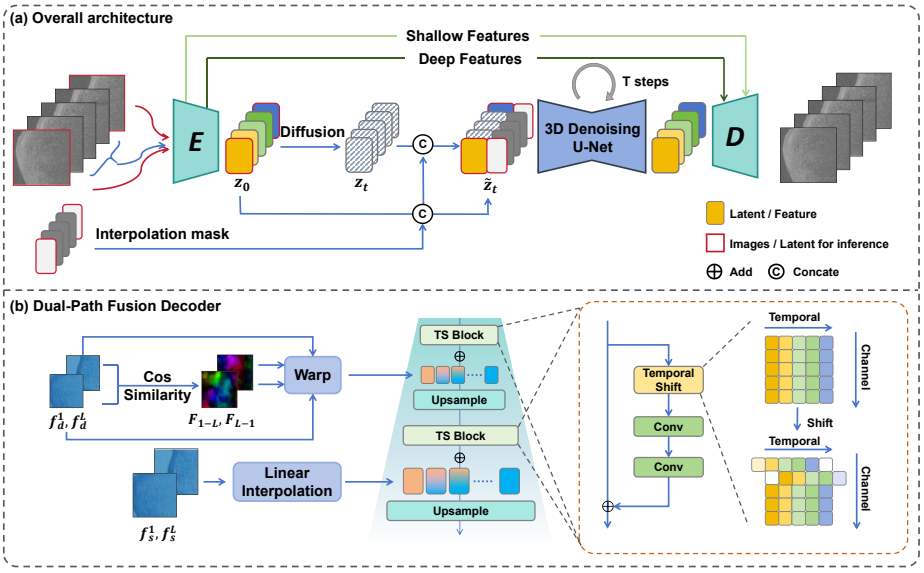

Fig. 2. (a) The overall architecture of *CardioInterp*. The first and final scans, together with an interpolation mask, are used as conditions for intermediate scanes synthesis. (b) An overview of the dual-path fusion decoder.

where x_t is the noisy image at timestep t, $\bar{\alpha}_t := \prod_{i=1}^{T} \alpha_i$, and α_i is hyper-parameters relevant to variance. Thus, x_t can be formulated as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}}\epsilon; \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

For the purpose of interpolation, intermediate latent codes have to be learned conditioned on the first and last ones. Considering the OCT latent code: $\mathbf{z}_t = \{\mathbf{z}_t^i\}_{i=1}^{L}$ where $\mathbf{z}_t^i \in \mathbb{R}^{h\times w\times c}$ and L is the total number of latent codes within the OCT latent. For each slice in this latent volume, a binary mask is concatenated along the channel dimension into the latent tensor, serving as conditional indicators to specify whether the latent code functions as an input constraint:

$$\tilde{\mathbf{z}}_t = \{\tilde{\mathbf{z}}_t^i = [\mathbf{z}_t^i, \mathbf{m}^i]\}_{i=1}^{L}\}_{i=1}^{L}, \tilde{\mathbf{z}}_t^i \in \mathbb{R}^{h\times w\times(c+1)}, \tag{3}$$

$$\tilde{\mathbf{z}}_0 = \{\tilde{\mathbf{z}}_0^i = [\mathbf{z}_0^i, \mathbf{m}^i]\}_{i=1}^{L}\}_{i=1}^{L}, \tilde{\mathbf{z}}_0^i \in \mathbb{R}^{h\times w\times(c+1)}, \tag{4}$$

$$\tilde{\mathbf{z}}_t \leftarrow \tilde{\mathbf{z}}_t \odot (1-\mathbf{m}) + \tilde{\mathbf{z}}_0 \odot \mathbf{m}, \tag{5}$$

where $\mathbf{m} = \{\mathbf{m}^i\}_{i=1}^{L}, \mathbf{m}^i \in \mathbb{R}^{h\times w\times 1}$ is the binary mask. To train the interpolation task, the first and last binary masks $\{\mathbf{m}^i\}_{i=1,L}^{2}$ are set to ones and others are set to zeros. Thus, the training objective for the conditional DDPM becomes:

$$\mathcal{L}_{\text{condition}}(\theta) := \|\boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_t, t) - \boldsymbol{\epsilon}\|_2^2. \tag{6}$$

Inference The first and last scans serve as inputs of the encoder, generating latent representation $\{z_0^i\}_{i=1}^2$. Pure noise $z_T \sim \mathcal{N}(0, \mathbf{I})$ is added between slices in z_0 to get $\{\mathbf{z}_t\}_{i=1}^L$. Interpolation mask is then concated $\tilde{\mathbf{z}}_t = [\mathbf{z_t}, \mathbf{m}]$ and finally iteratively denoised to generate interpolated latents.

2.2 Dual-Path Fusion Decoder

As shown in Fig. 2 (a), the dual-path fusion decoder integrates shallow-layer features and deep semantic features from the encoder. Through a dual-path fusion strategy, these hierarchical representations are fused within the decoder architecture via residual learning, ensuring structural continuity and fine detail retention in the synthesized scans. Furthermore, TSM [8] is employed as the P-3D Block to explicitly model inter-slice dependencies, thereby enhancing slice-wise consistency and coherence in the generated sequences without compromising computational efficiency.

Dual-Path Fusion Strategy As illustrated in Fig. 2 (b), to capture the continuous variations between deep semantic features, we employ a feature flow warping-based interpolation method for intermediate feature generation between spatially discontinuous leading and trailing frames. In natural video interpolation tasks, pre-trained motion estimation models are typically utilized to extract inter-frame motion cues for intermediate synthesis [17, 13, 10]. However, cardiovascular OCT imaging lacks domain-specific deformation estimation models, rendering conventional motion priors unreliable. To mitigate this, we utilize feature flow to estimate deformation flow among slices at the feature level [13], enforcing anatomical continuity constraints.

The bidirectional feature flow is derived through cosine-similarity-guided correspondence matching between deep semantic features extracted from the leading and trailing frames. For each spatial position (i, j) in feature f_d^1, the pixel in the other feature f_d^0 with the highest cosine similarity is selected as its corresponding location, then the feature flow is obtained:

$$F^{0\to1}(x, y) = \arg\max_{i,j} \langle f_d^0(x, y), f_d^1(i, j) \rangle, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. Similarly, the reverse feature flow $F^{1\to0}$ can also be obtained. We estimate the feature flow from intermediate time δ to time 0 as: $F^{\delta\to0}(x, y) = \delta F^{1\to0}(x, y)$, and feature flow from intermediate time δ to time 1 as: $F^{\delta\to1}(x, y) = (1 - \delta)F^{0\to1}(x, y)$. Upon knowing the two feature flow field, we can synthesize the intermediate feature $\hat{f}_d^\delta$ via a time-weighted interpolation:

$$\hat{f}_d^\delta = \delta \cdot g(f_d^0, F_{\delta\to0}) + (1 - \delta) \cdot g(f_d^1, F_{\delta\to1}), \tag{8}$$

where $g(\cdot, \cdot)$ is backward warping. Finally, the estimated intermediate features are added to features of decoder.

For the shallow layers, we implement linear interpolation to synthesize texture features between input frames. As shallow-layer features inherently exhibit

Table 1. Quantitative results of different medical slice interpolation methods.

Methods	×2		×4		×6	
	PSNR↑	SSIM(%)↑	PSNR↑	SSIM(%)↑	PSNR↑	SSIM(%)↑
I^3Net	29.18	56.50	**28.96**	53.97	-	-
TSCNet	10.53	11.38	14.20	21.33	-	-
FLAVR	29.09	51.18	28.60	47.78	28.07	47.56
Ours	**29.94**	**59.21**	28.78	**56.00**	**28.59**	**51.80**

higher spatial resolution and larger dimensionality, direct computation of feature flow estimation would substantially increase computational complexity and memory consumption. To address this challenge, we adapt linear interpolation specifically for shallow texture features, as shown in Eq. 9, to estimate the feature $\hat{f}_s^\delta$ at time δ:

$$\hat{f}_s^\delta = (1-\delta) \cdot f_s^0 + \delta \cdot f_s^1. \tag{9}$$

Temporal Shift for P-3D Decoding P-3D architectures typically employ 3D convolutions (e.g., Conv3D with kernel size 3×1×1) to explicitly model temporal dependencies in video sequences [18], which introduce extra computational overhead due to spatiotemporal feature aggregation across adjacent frames. To mitigate computational demands while preserving inter-slice coherence, our framework integrates TSMs to boost efficency and accuracy. This design effectively redistributes temporal information through channel-wise feature shifting operations, achieving deformation-aware feature propagation along the temporal axis without requiring explicit 3D convolutions.

3 Experiments and Results

3.1 Dataset

We curated the dataset that comprises real-world cardiovascular OCT sequences. It contains 387 clinical cases, with a total of 788 volumetric scans, among which, 730 volumes were retained after excluding samples with sequence length below 20 slices. This dataset comprises 645 volumes acquired using the Vivolight intravascular OCT system and 85 volumes using the Abbott intravascular OCT system, with each volume dimensioned at $400 \times 497 \times 1,025$ voxels. The B-slices were captured at 0.2 mm inter-slice intervals, covering a vascular scanning length of 80 mm. The A-lines were spaced at 7.47×10^{-3} mm intervals, resulting in an effective scanning depth of 7.66 mm.

3.2 Implementation Details

The dataset was partitioned into 14,224 contiguous sub-volumes of dimensions $20 \times 497 \times 1{,}025$, then split into training, validation, and test sets at an 8:1:1

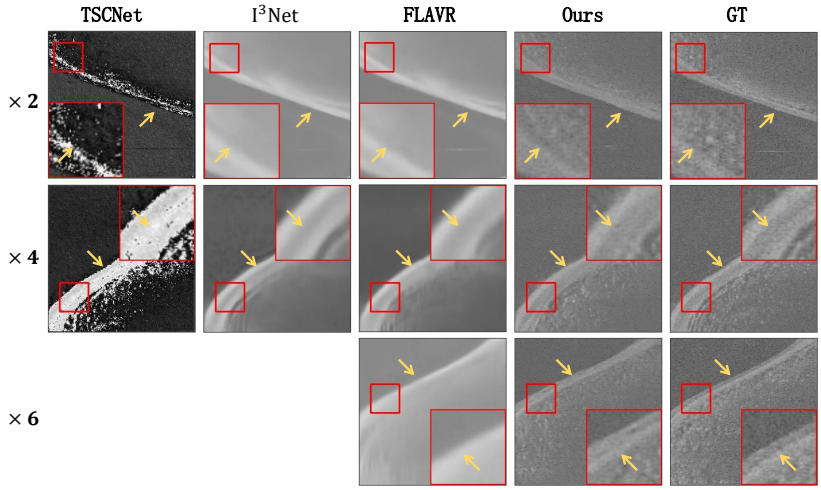

Fig. 3. Visual comparison of different medical slice interpolation methods on our cardiovascular OCT dataset.

ratio. For each sub-volume, we randomly selected a stride from $1, 2, 3$ slices and extracted $7 \times 256 \times 256$ voxel patch volumes from the selected sub-volume using the chosen stride with random positioning. Our implementation is primarily based on video generation model LVDM [3]. We first train the autoencoder with 100 epochs on learning rate (lr) $lr = 3.60 \times 10^{-5}$ and $bs = 2$. Then we train the DDPM process with 60 epochs on $lr = 8 \times 10^{-5}$ and batch size (bs) $bs = 10$. The training process is conducted on 4 NVIDIA A6000 GPUs. At inference, we adopt DDIM sampler and unconditional guidance.

3.3 Comparisons with SOTA and Baselines

In Table 1, we summarized PSNR and SSIM scores on cardiovascular OCT dataset under $\times 2$, $\times 4$ and $\times 6$ upscale. Two types of methods were compared: (1) medical slices interpolation methods, including I^{3}Net [14], TSCNet [9], and (2) kernel-based video interpolation method FLAVR [6]. I^{3}Net and TSCNet were trained under $\times 2$ upscale and evaluated for $\times 2$ and $\times 4$ upscales. FLAVR was trained and evaluated under $\times 2$, $\times 4$, and $\times 6$ upscales. For *CardioInterp*, it was trained under $\times 6$ and evaluated under $\times 2$, $\times 4$ and $\times 6$ upscales.

As shown in Table 1, *CardioInterp* achieved the best performance of PSNR 28.59 and SSIM 51.80% in $\times 6$ upscale tasks, outperforming FLAVR by margins of $+0.52$ dB and 4.24% in the respective metrics. Remarkably, *CardioInterp* demonstrated zero-shot generalization capability to $\times 2$ and $\times 4$ scaling factors despite being exclusively trained on $\times 6$ upsampling objectives. Through inherent spatial integration learned during high-ratio upscaling training process, *CardioInterp* achieved competitive PSNR and SSIM scores of $+0.76$ and 2.71% compared to SOTA models specifically optimized for $\times 2$ upscale tasks. Visual

Table 2. Ablation studies of different modules in *CardioInterp* on ×6 upscaling.

P3D	Dual-Path Addition	Dual-Path Fusion	PSNR↑	SSIM(%)↑
✗	✗	✗	27.91	46.13
✓	✗	✗	28.08	46.67
✓	✓	✗	28.24	49.07
✓	✓	✓	**28.59**	**51.80**

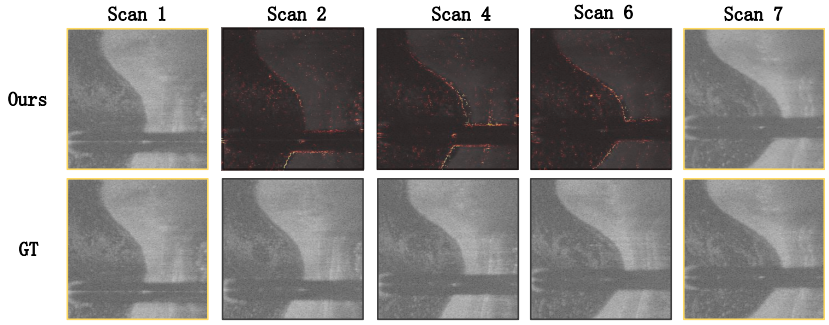

Fig. 4. Visual results used to demonstrate the continuity of generated slices.

comparisons in Fig. 3 show *CardioInterp* achieves enhanced of vascular structures in synthetic images, with higher clarity in critical anatomical features such as vessel wall boundaries and stratified intraluminal layers. Fig. 4 demonstrates the continuity of generated slices with visual output. As there is no sagittal view to prove structure continuity of generated B-slices, we put a sequence of slices and corresponding error maps to show the continuity of the vascular wall.

3.4 Ablation Studies

We conduct comprehensive ablation experiments to compare our proposed *CardioInterp* with different variants in Table 2. Note that dual-path addition denotes that dual path features are added directly into the decoder rather than fused using flow and linear interpolation. Quantitative results in Table 2 reveal that With the employment of the dual-path strategy, the continuity and fidelity of the generated results have been enhanced. Specifically, the PSNR has increased by +0.51, and the SSIM has increased by +5.13%.

4 Discussion and Conclusion

In this work, we present *CardioInterp*, the first model for cardiovascular OCT volume interpolation. *CardioInterp* integrates the diffusion process with expressive advanced spatiotemporal modeling, featuring a novel decoder that enhances structural continuity and imaging fidelity. Extensive experiments were conducted

to demonstrate its effectiveness, achieving the highest PSNR and SSIM scores compared with previous approaches. *CardioInterp* successfully synthesizes continuous high-fidelity OCT B-slices with preserved tissue microstructures under strong anisotropic acquisition patterns. However, the current implementation is limited to a reconstruction scale of 256×256 pixels, which is below the standard diagnostic-grade resolution of $497 \times 1,025$ pixels. Future research should address these interpolation scale constraints to maintain continuity and fidelity for accurate intravascular assessments.

Acknowledgments. This work is supported by the Science, Technology, and Innovation Commission (STIC) of Shenzhen Municipality (SGDX20220530111005039), the Research Grants Council (RGC) of Hong Kong SAR (GRF14216222, GRF14201824), the Innovation and Technology Fund (ITF) of Hong Kong SAR (ITS/252/23).

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

References

1. Bezerra, H.G., Costa, M.A., Guagliumi, G., Rollins, A.M., Simon, D.I.: Intracoronary optical coherence tomography: a comprehensive review: clinical and research applications. JACC: Cardiovascular Interventions **2**(11), 1035–1046 (2009)
2. Fang, C., Wang, L., Zhang, D., Xu, J., Yuan, Y., Han, J.: Incremental cross-view mutual distillation for self-supervised medical ct synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20677–20686 (2022)
3. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
5. Jain, S., Watson, D., Tabellion, E., Poole, B., Kontkanen, J., et al.: Video interpolation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7341–7351 (2024)
6. Kalluri, T., Pathak, D., Chandraker, M., Tran, D.: Flavr: Flow-agnostic video representations for fast frame interpolation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2071–2082 (2023)
7. Li, L., Qiu, J., Saha, A., Li, L., Li, P., He, M., Guo, Z., Yuan, W.: Artificial intelligence for biomedical video generation. arXiv preprint arXiv:2411.07619 (2024)
8. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019)
9. Lu, Z., Li, Z., Wang, J., Shi, J., Shen, D.: Two-stage self-supervised cycle-consistency network for reconstruction of thin-slice mr images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 3–12. Springer (2021)
10. Lyu, Z., Li, M., Jiao, J., Chen, C.: Frame interpolation with consecutive brownian bridge diffusion. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3449–3458 (2024)

11. Peng, C., Lin, W.A., Liao, H., Chellappa, R., Zhou, S.K.: Saint: spatially aware interpolation network for medical slice synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7750–7759 (2020)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
13. Shen, L., Liu, T., Sun, H., Ye, X., Li, B., Zhang, J., Cao, Z.: Dreammover: Leveraging the prior of diffusion models for image interpolation with large motion. In: European Conference on Computer Vision. pp. 336–353. Springer (2024)
14. Song, H., Mao, X., Yu, J., Li, Q., Wang, Y.: I 3 net: Inter-intra-slice interpolation network for medical slice synthesis. IEEE Transactions on Medical Imaging (2024)
15. Tearney, G.J., Brezinski, M.E., Bouma, B.E., Boppart, S.A., Pitris, C., Southern, J.F., Fujimoto, J.G.: In vivo endoscopic optical biopsy with optical coherence tomography. Science **276**(5321), 2037–2039 (1997)
16. Ughi, G.J., Adriaenssens, T., Larsson, M., Dubois, C., Sinnaeve, P.R., Coosemans, M., Desmet, W., D'hooge, J.: Automatic three-dimensional registration of intravascular optical coherence tomography images. Journal of biomedical optics **17**(2), 026005–026005 (2012)
17. Wu, G., Tao, X., Li, C., Wang, W., Liu, X., Zheng, Q.: Perception-oriented video frame interpolation via asymmetric blending. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2753–2762 (2024)
18. Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., Wong, T.T.: Tooncrafter: Generative cartoon interpolation. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)