

EchoTracker2: Enhancing Myocardial Point Tracking by Modeling Local Motion

Md Abulkalam Azad^{1,2}[0009–0006–7177–4961], Vegard Holmstrøm¹, John Nyberg¹, Lasse Lovstakken¹, Håvard Dalen^{1,2}, Bjørnar Grenne^{1,2}, and Andreas Østvik^{1,2,3}[0000–0003–3895–2683]

¹ Norwegian University of Science and Technology, Trondheim, Norway

² Clinic of Cardiology, St. Olavs Hospital, Trondheim, Norway

³ SINTEF Digital, Trondheim, Norway

{md.a.azad,andreas.ostvik}@ntnu.no

Abstract. Myocardial point tracking (MPT) has recently emerged as a promising direction for motion estimation in echocardiography, driven by advances in general-purpose point tracking methods. However, myocardial motion fundamentally differs from motion encountered in natural videos, as it arises from physiologically constrained deformation that is spatially and temporally continuous throughout the cardiac cycle. Consequently, motion trajectories typically remain locally confined despite substantial tissue deformation. Motivated by these properties, we revisit the architectural design for MPT and find that coarse initialization in commonly used two-stage coarse-to-fine architectures may be unnecessary in this domain. In this work, we propose a fine-stage-only architecture, **EchoTracker2**, which enriches pixel-precise features with local spatiotemporal context and integrates them with long-range joint temporal reasoning for robust tracking. Experimental results across in-distribution, out-of-distribution (OOD), and public synthetic datasets show that our model improves position accuracy by 6.5% and reduces median trajectory error by 12.2% relative to a domain-specific state-of-the-art (SOTA) model. Compared to the best general-purpose point tracking method, the improvements are 2.0% and 5.3%, respectively. Moreover, EchoTracker2 shows better agreement with expert-derived global longitudinal strain (GLS) and enhances test-rest reproducibility. Source code is available at: <https://github.com/riponazad/ptecho>.

Keywords: Point tracking · Motion estimation · Echocardiography · Ultrasound · Strain imaging.

1 Introduction

Speckle tracking echocardiography is a well-established approach for cardiac motion estimation and is clinically used to quantify strain, a sensitive marker of cardiac function and abnormal tissue deformation [11, 14, 15, 17]. Tracking in general computer vision tasks must handle substantial variability in motion patterns arising from object movement, camera motion, occlusions and scene

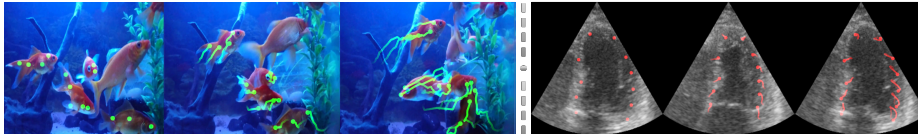


Fig. 1. Motion trajectories of selected query points are shown on the first, middle, and last frames for a natural video (left, TAP-Vid DAVIS [8]) and a cardiac video (right).

changes. Motion can be rigid or non-rigid, temporally inconsistent, making long-range correspondence inherently ambiguous. In contrast, myocardial motion in echocardiography is governed by physiological deformation that is spatially and temporally coherent, continuous, and constrained by anatomical structure [3]. Although ultrasound imaging presents challenges such as speckle noise, low contrast, and view-dependent artifacts, tissue motion itself follows smooth and cyclic trajectories driven by cardiac contraction and relaxation [7].

As in Fig. 1, queried points in natural videos may move abruptly in arbitrary directions with large displacements, whereas motion in echocardiographic sequences is comparatively structured and remains confined to a local spatial neighborhood across frames. Importantly, this behavior arises from biomechanical constraints of myocardial tissue rather than acquisition conditions alone [6]. Continuity of deformation across neighboring regions and periodic contraction-relaxation dynamics enforce strong spatial and temporal priors on motion trajectories throughout the cardiac cycle [19]. Consequently, while general-purpose tracking methods must accommodate unconstrained global motion, myocardial motion estimation can instead exploit these structural constraints. Motivated by this, we hypothesize that focusing solely on local motion is sufficient for robust tracking in echocardiography, while enabling improved pixel-level accuracy that is critical for downstream functional measurements such as strain.

In this work, we propose **EchoTracker2**, a single-stage architecture for fine-grained myocardial point tracking across a full cardiac cycle that explicitly leverages locally constrained motion. In contrast to existing coarse-to-fine approaches, such as EchoTracker [1] and MyoTracker [4], our method finds correspondence within a confined spatial neighborhood and integrates local spatiotemporal feature representations with joint temporal reasoning across neighboring trajectories. Experimental results demonstrate improved tracking performance and generalizability across in-distribution, OOD, and public synthetic datasets. Furthermore, EchoTracker2 improves agreement and reproducibility in downstream GLS measurements compared to existing methods.

2 Methods

Given a cardiac video $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times C}$ and a set of N query points $\mathcal{P} = \{(x_i^{t_q}, y_i^{t_q}) \in \mathbb{R}^2 \mid i = 0, \dots, N-1\}$ selected on the myocardium from an arbitrary frame $t_q \in \{0, \dots, T-1\}$, the objective is to estimate the point trajectories $\mathcal{T} = \{(x_i^t, y_i^t) \mid i = 0, \dots, N-1; t = 0 \dots, T-1\}$ throughout the cardiac cycle.

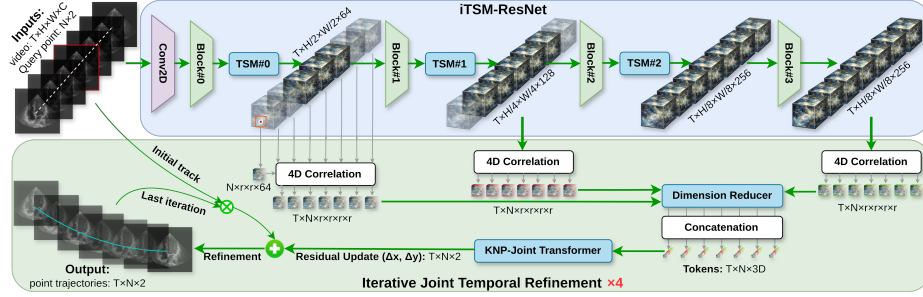


Fig. 2. The architecture of EchoTracker2 and the computational flow for tracking a single point across ultrasound frames. Within each TSM block, temporal context is aggregated from adjacent frames, progressively expanding the temporal receptive field toward deeper blocks (indicated by the transparency of feature volumes). Although shown only for the middle frame (red border), the same process applies to all frames. The final trajectories are obtained through four iterations of joint temporal refinements.

2.1 EchoTracker2

We design EchoTracker2 to model the locally constrained deformation that characterizes myocardial motion. Unlike general-purpose tracking methods, which rely on coarse initialization for global motion to accommodate large and unpredictable displacements, myocardial trajectories are governed by continuous tissue deformation and exhibit limited inter-frame displacement. This allows correspondence estimation to be performed within a confined spatial neighborhood without requiring global correlation. Hence, we eliminate the commonly used coarse initialization, focus on learning pixel-precise local spatiotemporal representations, and refine trajectories through joint temporal reasoning across neighboring points. An overview of EchoTracker2 is shown in Fig. 2, and its components are described below.

Feature extraction. Following Doersch et al. [8, 9], we adopt the temporal shift module (TSM)-ResNet [13] to extract temporally-aware features. In prior work, TSMs are applied only to the deepest ResNet block, despite multi-scale correlation being computed across features from all blocks. This introduces a mismatch between spatial resolution and temporal context, as shallower feature maps used for correspondence estimation remain temporally independent. In the context of myocardial motion, where correspondence is estimated within a confined spatial neighborhood and trajectories evolve smoothly over time, feature representations at all scales should encode local temporal information. We therefore integrate lightweight TSM layers between each ResNet block, yielding a temporally-enhanced backbone referred to as **iTSM-ResNet**. Each TSM aggregates context from adjacent frames, progressively expanding the temporal receptive field to 7 frames (3 backward and 3 forward) at the deepest block, as illustrated in Fig. 2. Let $\mathcal{F}^{(0)} \in \mathbb{R}^{T \times \frac{H}{k^{(0)}} \times \frac{W}{k^{(0)}} \times d^{(0)}}$ denote the intermediate frame features before the first ($\ell = 0$) ResNet block $\mathcal{B}_0$, where $k^{(0)}$ is the spatial

downsampling factor and $d^{(0)}$ is the feature dimension. The feature extraction process of iTSM-ResNet can then be summarized as;

$$\mathcal{F}^{(\ell+1)} = \begin{cases} \mathcal{TSM}_\ell(\mathcal{B}_\ell(\mathcal{F}^{(\ell)})), & \ell \in \{0, 1, 2\} \\ \mathcal{B}_\ell(\mathcal{F}^{(\ell)}), & \ell = 3 \end{cases} \quad (1)$$

Correlation computation. We estimate correspondence by computing feature correlation within a local spatial neighborhood. While 2D correlation is widely used, we adopt 4D correlation [5] due to its robustness to matching ambiguity, a key advantage for ultrasound imaging. As illustrated for a single point at one feature level in Fig. 2, query features $\mathcal{Q}^\ell \in \mathbb{R}^{N \times r \times r \times d^{(\ell)}}$ are bilinearly sampled for N points from the query frame within an $(r \times r)$ window at level ℓ . Corresponding frame features $\mathcal{F}_c^{(\ell)} \in \mathbb{R}^{T \times N \times r \times r \times d^{(\ell)}}$ are sampled at the same spatial resolution $(r \times r)$ along the previously estimated trajectories (in the first iteration, point locations from the query frame are propagated to all frames). All-pair cosine similarity is computed between $\mathcal{Q}^{(\ell)}$ and $\mathcal{F}_c^{(\ell)}$ to obtain high-dimensional correlation maps $\mathcal{C}^{(\ell)} \in \mathbb{R}^{T \times N \times r \times r \times r \times r}$. These maps are subsequently encoded into a compact representation of dimensionality D using a learnable encoder [5] and concatenated across three feature scales $(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}, \mathcal{F}^{(4)})$, yielding multi-scale correlation tokens $\mathcal{C} \in \mathbb{R}^{T \times N \times 3D}$.

Joint Temporal refinement. Although features enriched with local spatiotemporal context are often sufficient for short-term tracking, the accuracy may degrade in the presence of out-of-plane motion or image artifacts. Besides, myocardial deformation exhibits spatially coherent motion among neighboring regions, and the coherency differs progressively for distant regions. Thus, nearby points have stronger motion dependency than distant ones. To account for this property, we extend the transformer of Cho et al. [5] by applying multi-head self-attention [16] both along trajectories and across K nearest points simultaneously for joint temporal reasoning, referred to as the **KNP-Joint** transformer (Fig. 2). The module takes the correlation tokens $\mathcal{C}$ as input and predicts residual trajectory updates $(\Delta x, \Delta y) \in \mathbb{R}^{T \times N \times 2}$. The final refined trajectories are obtained iteratively as $\mathcal{T}^{(m)} = \mathcal{T}^{(m-1)} + (\Delta x, \Delta y)^{(m)}$ after $m = 4$ iterations.

3 Experiments

3.1 Datasets

We use real-world in-house echocardiography datasets acquired using GE Vivid E95 and E9 systems, under approval from the regional ethics committee and with informed patient consent (Table 1). Each video covers a full cardiac cycle from end-diastole to the subsequent end-diastole frame. For each sequence, sparse myocardial point trajectories were generated along the left or right ventricular

Table 1. Overview of the echocardiographic datasets.

Dataset	Patients	Videos	Pathology	Views
Training	643	1922	Healthy	A4C · A2C · A3C
In-distribution	371	1112	Mixed	A4C · A2C · A3C
OOD RV	1254	1331	Healthy	A4C (RV-focused)
CAMUS	98	196	Undefined	A4C
Agreement	203	609	Mixed	A4C · A2C · A3C
Reproducibility	127	381	Mixed	A4C · A2C · A3C

Mixed: Healthy subjects and patients with heart failure, myocardial infarction or atrial fibrillation.
A4C: Apical four-chamber, A2C: Apical two-chamber, A3C: Apical three-chamber/long-axis

wall (LV or RV) using a proprietary semi-automatic algorithm from GE Health-Care. In brief, the ventricular wall was initialized in a mid-systolic frame, after which the software propagated N myocardial points throughout the cardiac cycle to produce frame-wise coordinate trajectories. These resulting trajectories were refined and quality-controlled by experienced cardiologists (>20 years). During this process, tracking failures, anatomically implausible trajectories, and incorrectly placed points were manually corrected or excluded according to image quality and tracking reliability. The final training labels therefore consisted of verified coordinate trajectories for N sparse myocardial points across all frames of the cardiac cycle. On average, each video contained 50-80 tracked myocardial points over 50-150 frames per cardiac cycle.

All models, including SOTA baselines and EchoTracker2, are trained or fine-tuned exclusively on the Training dataset and evaluated on separate in-house and public datasets. The In-distribution dataset contains LV motion from both healthy subjects and patients with cardiac pathology acquired from standard views. The OOD RV dataset consists of RV-focused acquisitions from healthy subjects and represents a distribution shift in both anatomy and imaging view. We additionally evaluate generalizability on the publicly available synthetic CAMUS dataset [10], which provides myocardial trajectories sampled across multiple endocardial-to-epicardial contours, in contrast to the single mid-wall trajectory used during training. Finally, the Agreement and Reproducibility datasets are used to assess GLS agreement with experts and test-retest stability, respectively. The latter consists of repeated acquisitions from the same patient obtained within the same session under comparable physiological conditions.

3.2 Evaluation Metrics

Tracking performance is commonly evaluated as the average percentage of tracked points that remain within a spatial tolerance x of their ground-truth locations along the trajectories, averaged over multiple pixel thresholds ($<\delta_{avg}^x$) [8, 9]. Standard protocols include tolerances up to $x = 16$ pixels, however, such thresholds are less relevant for clinical strain estimation where small tracking errors may substantially affect deformation measurements. We therefore report $<\delta_{avg}^x$ only for $x \in \{1, 2, 4\}$ pixels to better reflect clinically meaningful tracking accuracy. In addition, we report the median trajectory error (MTE) in pixels and

the average inference time (AIT) in seconds per video. Distances are computed using the $\mathcal{L}_1$ distance at a spatial resolution of 256×256 pixels.

Clinical evaluation is performed using peak GLS, which quantifies the relative change of the longitudinal ventricular length from end-diastole to minimum ventricular length. Agreement with expert-derived measurements is assessed using the mean and standard deviation of differences, while reproducibility is evaluated using the mean absolute difference (MAD) and coefficient of variation (CV).

3.3 Implementation details

All SOTA models (based on their open-source implementations) and EchoTracker2 are implemented in PyTorch. For training and fine-tuning, we adopt the impartial motion augmentation strategy proposed in [2]. Each model is trained for 100 epochs (8,620 augmented clips, each containing 36 frames with 40 query points; batch size 4). Training each model requires approximately one week on a single A6000 (48GB) GPU. The training hyperparameters (learning rate = 5×10^{-4} , one-cycle scheduler, and AdamW optimizer) are kept identical across models to ensure fair comparison. All evaluations in Section 4 are performed on a single RTX 3090 (24GB) GPU. For supervision, we compute the $\mathcal{L}_1$ distance between the reference trajectories $\mathcal{T}^{gt}$ and the estimated trajectories $\mathcal{T}$, and apply exponential weighting ($\gamma = 0.8$) over the m refinement iterations as

$$\mathcal{Loss} = \frac{1}{m} \sum_{i=1}^m \gamma^{m-i} \mathcal{L}_1(\mathcal{T}^{(i)}, \mathcal{T}^{gt}) \quad (2)$$

3.4 Ablation Study

We conduct ablation studies to evaluate the impact of key architectural design choices. All variants are trained for 50 epochs on 200 sequences (36 frames each) using identical hyperparameters and validated on a held-out subset of 40 patients from the In-distribution dataset to ensure consistent comparison.

Local Spatial Correlation. Since EchoTracker2 restricts correspondence estimation to a local spatial neighborhood, the correlation window must be sufficiently large to capture typical inter-frame displacements. Although several SOTA methods [1, 5, 12] have used a 7×7 window, the optimal size depends on input resolution, downsampling factor, and motion characteristics of the imaging domain. We therefore evaluate four window sizes for 4D correlation, as shown in Table 2.a. A window size of 9×9 provides the best trade-off between tracking accuracy and computational cost for EchoTracker2.

Local Temporal Feature Enrichment. We explore several strategies to enrich frame features with neighboring temporal context (Table 2.b). An efficient approach is to fuse precomputed local spatial features from adjacent frames prior

Table 2. Ablation studies. Impact of (a) different spatial window sizes for 4D correlation, (b) local temporal feature enrichment, (c) joint temporal reasoning.

Configuration	In-distribution					OOD RV				
	$<\delta^1\uparrow$	$<\delta^2\uparrow$	$<\delta^4\uparrow$	$MTE\downarrow$	$AIT\downarrow$	$<\delta^1\uparrow$	$<\delta^2\uparrow$	$<\delta^4\uparrow$	$MTE\downarrow$	$AIT\downarrow$
<i>(a) Spatial window size</i>										
5×5	29.99	54.56	80.78	2.24	0.17	20.65	38.97	62.59	3.73	0.19
7×7	30.53	54.65	80.93	2.22	0.21	21.42	39.92	63.38	3.70	0.21
9×9	31.82	55.83	81.42	2.17	0.27	22.34	41.48	65.01	3.54	0.28
11×11	30.88	55.04	81.15	2.21	0.35	21.04	39.94	63.98	3.61	0.36
<i>(b) Temporal feature enrichment</i>										
BL+Fusion(add)	32.44	56.40	81.57	2.14	0.27	22.40	40.75	63.68	3.59	0.29
BL+Fusion (cat)	32.82	57.01	82.52	2.10	0.27	22.83	41.91	65.65	3.44	0.30
bTSM-ResNet	32.98	57.14	82.50	2.10	0.33	23.29	42.75	66.24	3.45	0.37
iTSM-ResNet	33.31	57.40	82.46	2.08	0.32	23.20	42.27	65.85	3.43	0.36
<i>(c) Joint reasoning strategy</i>										
Joint-Reasoning	33.12	57.29	82.39	2.09	0.34	23.16	42.19	65.51	3.50	0.37
Cross-Attention	32.70	56.56	81.76	2.13	0.29	23.05	42.13	65.32	3.49	0.32
KNP-Joint	33.31	57.40	82.46	2.08	0.32	23.20	42.27	65.85	3.43	0.36

to correlation, either through addition or concatenation, similar to [1, 18]. Alternatively, temporal reasoning can be incorporated directly into the backbone using learnable modules such as TSM-ResNet [9]. We implement TSM in two ways: blockwise (bTSM-ResNet), where features from each block pass through a single-layer TSM without altering the original data flow, and injection (iTSM-ResNet), where a TSM is inserted after each block so its output is propagated to the subsequent block.

Long-range Temporal Reasoning. We first extend the transformer of Cho et al. [5] with self-attention across all points for full joint reasoning and compare it to the cross-attention-based transformer of Karaev et al. [12]. Our final variant, KNP-Joint temporal transformer, performs joint reasoning over the $k = 10$ nearest neighboring trajectories and achieves the best performance (Table 2.c).

4 Results

Local Motion in Echocardiography. Table 3 shows that removing the initialization stage from EchoTracker and LocoTrack results in negligible differences in tracking performance, despite both models being trained with it. This suggests that modeling local motion alone is sufficient and reduces computational cost.

Tracking performance. To benchmark EchoTracker2, we compare its performance with leading SOTA models from both the general-purpose and echocardiography domains (Table 4). Although pretrained PIPs++, LocoTrack (base), and CoTracker3 (offline) were trained on large-scale general video datasets [5,

Table 3. Performance comparison of tracking and computational cost without initialization stage on In-distribution dataset.

Initial Stage?	EchoTracker [1]					LocoTrack [5]				
	$<\delta^1\uparrow$	$<\delta^2\uparrow$	$<\delta^4\uparrow$	$MTE\downarrow$	$AIT\downarrow$	$<\delta^1\uparrow$	$<\delta^2\uparrow$	$<\delta^4\uparrow$	$MTE\downarrow$	$AIT\downarrow$
Yes	37.07	64.01	88.73	1.82	0.15	39.95	65.70	89.19	1.73	0.23
No	37.54	63.89	88.30	1.83	0.12	39.55	65.40	89.23	1.73	N/A*

*The initial stage could not be fully removed due to architectural dependencies, although it was ignored during tracking. Consequently, a meaningful AIT could not be obtained.

Table 4. Overall tracking performance of the different models.

Methods	In-distribution			OOD RV			CAMUS [10]		
	$<\delta_{avg}^x\uparrow$	$MTE\downarrow$	$AIT\downarrow$	$<\delta_{avg}^x\uparrow$	$MTE\downarrow$	$AIT\downarrow$	$<\delta^x\uparrow$	$MTE\downarrow$	$AIT\downarrow$
PIPs++ [18]	57.40	2.06	0.14	41.78	3.40	0.16	54.90	2.25	0.17
SpeckNet [2]	59.56	1.97	0.20	43.66	3.53	0.26	64.24	1.81	0.26
EchoTracker [1]	63.27	1.82	0.15	46.63	3.17	0.19	70.33	1.58	0.18
CoTracker3 [12]	64.62	1.74	0.55	47.60	3.02	0.64	71.96	1.43	0.70
LocoTrack [5]	64.95	1.73	0.23	49.34	2.93	0.24	73.71	1.43	0.29
EchoTracker2(ours)	66.94	1.63	0.33	49.84	2.84	0.37	75.17	1.33	0.47

12, 18] and further finetuned on our training dataset, EchoTracker2, trained only on echocardiography videos for 100 epochs, outperforms them across In-distribution, OOD RV, and synthetic CAMUS datasets. Averaged across datasets, EchoTracker2 improves position accuracy by 6.52% and reduces MTE by 12.22% relative to EchoTracker, and surpasses the best performing general-purpose point tracking model, LocoTrack, by 2.02% and 5.28%, respectively. Some qualitative tracking results are attached in the supplement.

Evaluation of GLS agreement and test–retest reproducibility. Table 5 summarizes the GLS obtained using EchoTracker2 in comparison with expert observers, the best-performing general-purpose tracking model, and the domain-specific EchoTracker baseline. Agreement metrics reflect AI-to-observer variability, while the expert observer row reports inter-observer variability. Reproducibility is evaluated in a test–retest setting, with the expert observer row indicating intra-observer variability. EchoTracker2 demonstrates improved agreement with expert measurements compared to existing tracking methods, while also reducing the coefficient of variation in test–retest reproducibility.

Table 5. GLS agreement with expert observers and test-retest reproducibility.

Method	Agreement				Reproducibility		
	Mean $\pm$ SD	$\mu\downarrow$	$\sigma\downarrow$	MAD $\downarrow$	Mean $\pm$ SD	MAD $\downarrow$	CV $\downarrow$
Expert observers	−17.7 $\pm$ 3.4	0.58	1.80	1.44	−17.3 $\pm$ 3.5	1.29	6.8
EchoTracker	−15.9 $\pm$ 3.2	1.72	1.47	1.95	−15.5 $\pm$ 3.3	1.21	6.9
LocoTrack	−16.0 $\pm$ 3.2	1.60	1.46	1.86	−15.7 $\pm$ 3.3	1.30	7.2
EchoTracker2 (ours)	−16.9 $\pm$ 3.3	0.76	1.44	1.41	−16.5 $\pm$ 3.4	1.24	6.6

μ : mean difference (%), σ : standard deviation of differences (%), MAD: mean absolute difference (%), CV: coefficient of variation (%). Agreement columns show AI-vs-observer metrics; expert observer row shows inter-observer variability. Reproducibility expert row shows intraobserver.

5 Conclusion

We present **EchoTracker2**, a fine-stage-only MPT model, that enriches features with local spatiotemporal context and leverages 4D correlation for precise tracking. Our joint-temporal refinement across neighboring points improves robustness to out-of-plane motion and ultrasound artifacts. EchoTracker2 outperforms existing SOTA methods and demonstrates improved generalizability across in-distribution, OOD and public synthetic data. Furthermore, it improves agreement with expert-derived GLS measurements, while maintaining strong test-retest reproducibility. These results suggest that domain-specific motion modeling may enhance the reliability of deformation-based myocardial function assessment in routine echocardiographic practice.

Acknowledgments. This work was funded by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (2023-34187, 2024-37009).

Disclosure of Interests. J. Nyberg, H. Dalen, L. Løvstakken, B. Grenne, and A. Østvik are affiliated with ProCardio, a Centre for Research-based Innovation funded by the Research Council of Norway, in which GE HealthCare is a consortium partner. L. Løvstakken also serves as a part-time consultant for GE HealthCare.

References

1. Azad, M.A., Chernyshov, A., Nyberg, J., Tveten, I., Lovstakken, L., Dalen, H., Grenne, B., Østvik, A.: Echotracker: Advancing myocardial point tracking in echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 645–655. Springer (2024)
2. Azad, M.A., Nyberg, J., Dalen, H., Grenne, B., Lovstakken, L., Østvik, A.: Taming modern point tracking for speckle tracking echocardiography via impartial motion. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (IC-CVW). pp. 1115–1124 (2025)
3. Buckberg, G., Mahajan, A., Saleh, S., Hoffman, J.I., Coghlan, C.: Structure and function relationships of the helical ventricular myocardial band. *The Journal of thoracic and cardiovascular surgery* **136**(3), 578–589 (2008)
4. Chernyshov, A., Nyberg, J., Holmstrøm, V., Azad, M.A., Grenne, B., Dalen, H., Aase, S.A., Lovstakken, L., Østvik, A.: Low complexity point tracking of the myocardium in 2d echocardiography. *IEEE Access* **13**, 186992–187004 (2025)
5. Cho, S., Huang, J., Nam, J., An, H., Kim, S., Lee, J.Y.: Local all-pair correspondence for point tracking. In: European conference on computer vision. pp. 306–325. Springer (2024)
6. Claus, P., Omar, A.M.S., Pedrizzetti, G., Sengupta, P.P., Nagel, E.: Tissue tracking technology for assessing cardiac mechanics: principles, normal values, and clinical applications. *JACC: Cardiovascular Imaging* **8**(12), 1444–1460 (2015)
7. D’hooge, J.: Principles and different techniques for speckle tracking. *Myocardial imaging: tissue Doppler and speckle tracking* pp. 17–25 (2007)

8. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 13610–13626 (2022)
9. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 10061–10072 (2023)
10. Evain, E., Sun, Y., Faraz, K., Garcia, D., Saloux, E., Gerber, B.L., De Craene, M., Bernard, O.: Motion estimation by deep learning in 2d echocardiography: synthetic dataset and validation. *IEEE transactions on medical imaging* **41**(8), 1911–1924 (2022)
11. Gherbesi, E., Gianstefani, S., Angeli, F., Ryabenko, K., Bergamaschi, L., Armilotta, M., Guerra, E., Tuttolomondo, D., Gaibazzi, N., Squeri, A., et al.: Myocardial strain of the left ventricle by speckle tracking echocardiography: From physics to clinical practice. *Echocardiography* **41**(1), e15753 (2024)
12. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Co-tracker3: Simpler and better point tracking by pseudo-labelling real videos. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 6013–6022 (2025)
13. 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)
14. Mihos, C.G., Liu, J.E., Anderson, K.M., Pernetz, M.A., O’Driscoll, J.M., Aurigemma, G.P., Ujueta, F., Wessly, P., on Peripheral Vascular Disease; Council on Cardiovascular, A.H.A.C., Nursing; S., on Clinical Cardiology, C.: Speckle-tracking strain echocardiography for the assessment of left ventricular structure and function: a scientific statement from the american heart association. *Circulation* **152**(10), e96–e109 (2025)
15. Thomas, J.D., Edvardsen, T., Abraham, T., Appadurai, V., Badano, L., Banchs, J., Cho, G.Y., Cosyns, B., Delgado, V., Donal, E., et al.: Clinical applications of strain echocardiography: a clinical consensus statement from the american society of echocardiography developed in collaboration with the european association of cardiovascular imaging of the european society of cardiology. *Journal of the American Society of Echocardiography* (2025)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
17. Voigt, J.U., Pedrizzetti, G., Lysyansky, P., Marwick, T.H., Houle, H., Baumann, R., Pedri, S., Ito, Y., Abe, Y., Metz, S., Song, J.H., Hamilton, J., Sengupta, P.P., Kolias, T.J., d’Hooge, J., Aurigemma, G.P., Thomas, J.D., Badano, L.P.: Definitions for a common standard for 2d speckle tracking echocardiography: Consensus document of the eacvi/ase/industry task force to standardize deformation imaging. *Journal of the American Society of Echocardiography* **28**(2), 183–193 (2015)
18. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 19855–19865 (2023)
19. Zhu, Y., Papademetris, X., Sinusas, A.J., Duncan, J.S.: A coupled deformable model for tracking myocardial borders from real-time echocardiography using an incompressibility constraint. *Medical image analysis* **14**(3), 429–448 (2010)