Semi-supervised Key-Point Estimation for Echocardiography Video

Seok-Hwan Oh[1,*] Hyeon-Jik Lee[2,*] Guil Jung[2] Myeong-Gee Kim[1]
Young-Min Kim[2] Hyuksool Kwon[3] Hyeon-Min Bae[2,†]

[1]Barreleye Inc., Korea
[2]Department of Electrical Engineering, KAIST, Daejeon, Korea
[3]Department of Emergency Medicine, Seoul National University Bundang Hospital, Korea

Abstract

Echocardiography, a widely used imaging modality, offers real-time assessments of cardiac morphology and function, with a particular emphasis on left ventricular dynamics. Despite its clinical importance, existing automated methods for echocardiographic analysis struggle to ensure temporal consistency in left ventricular key-point trajectories, largely due to their reliance on static frame annotations. To overcome these challenges, we propose a semi-supervised trajectory refinement framework that employs inter-frame correlations to enhance key-point trajectory estimation across echocardiography videos. A semi-supervised trajectory learning scheme is presented to improve the efficacy of key-point trajectory analysis using unannotated echocardiography videos. The experiments present considerable improvements in both spatial accuracy and temporal stability of the left ventricle key-point trajectories, outperforming state-of-the-art baselines and demonstrating the clinical applicability for robust echocardiography analysis.

1. Introduction

Medical ultrasound is among the most widely adopted imaging modalities due to its cost-effectiveness, non-ionizing radiation, and real-time imaging capabilities. Of its various clinical applications, echocardiography (Echo) is particularly prominent because the Echo captures dynamic, real-time images of the beating heart, making it essential for diagnosing and monitoring various cardiac conditions [18, 20].

The heart is a dynamic organ that continuously adapts and remodels in response to physiological stresses and external perturbations. Intrinsic cardiac disorders and systemic insults lead to morphological abnormalities in the left ventricular (LV) wall and blood pool throughout the cardiac cycle [18]. Therefore, accurate assessment of LV functionality is significant for diagnosing major cardiac diseases and predicting clinical outcomes. For example, heart failure and cardiomyopathy cause significant changes in LV blood pool size. Variations in LV wall thickness during systole and diastole provide clinically significant biomarkers such as post-systolic thickening and delayed time-to-peak thickening in ischemia, as well as temporal thickening patterns observed during stress Echo. [2, 32].

However, it remains challenging to accurately diagnose LV disorders based solely on human observation. As a result, computer-aided quantification methods have been increasingly employed to facilitate LV functionality assessment [1, 33]. Traditionally, such measurements rely on manual image interpretation, where medical experts annotate key anatomical landmarks to estimate LV wall thickness and blood pool size. Such manual measurements are time-consuming, labor-intensive, and susceptible to considerable intra- and inter-operator variability [44]. Consequently, there is growing interest in automated Echo techniques that can provide more consistent, efficient, and objective measurements of LV morphology and function.

Recent studies have demonstrated that learning-based Echo analysis provides clinical phenotypes with greater accuracy compared to conventional image interpretation methods. In particular, the detection of LV key-points from clinically informative parasternal long-axis (PLAX) Echo videos has gained considerable attention. The methods enable the automated estimation of key clinical measurements, such as left ventricular internal dimensions (LVID), interventricular septum thickness (IVS), and left ventricular posterior wall thickness (LVPW) [9, 12, 21, 30].

Despite recent advancements, current Echo video analysis techniques demonstrates limited precision, particularly for more sophisticated cardiac disease analyses. The limitation arises from the fact that existing methodologies are mainly trained on datasets composed of static echocardio-

*These authors contributed equally to this work.
†Corresponding author: Hyeon-Min Bae (hmbae@kaist.ac.kr).

graphic images with single-frame key-point annotations, rather than full-video annotations [9]. Acquiring comprehensive key-point annotations across all frames of an Echo video presents a significant challenge, as it places a considerable workload on medical experts. As a result, publicly available datasets typically include only one or two annotated frames per video. Such lack of data hinders the ability to interpret inter-frame correlations necessary for maintaining the temporal stability of cardiac key-points and accurately identifying clinically relevant biomarkers.

To address the challenge, we propose a semi-supervised temporal key-point trajectory refinement framework, that enhances spatial precision and temporal stability of Echo video key-point trajectory analysis. Inspired by recent advances in point-tracking methodologies, we interpret the inter-frame relationships of the key-points and utilize the interdependencies to refine and stabilize key-point trajectories. In addition, for rigorous evaluation, we introduce, to the best of our knowledge, the first publicly available PLAX-view Echo video key-point dataset.

2. Related work

2.1. PLAX-view Echo

The PLAX-view Echo is a fundamental imaging method, that provides comprehensive insights into the structure and function of the LV. The PLAX-view Echo provides essential anatomical landmarks, including LVID, IVS, and LVPW, which are significant for the diagnosis and monitoring of various cardiac conditions [19, 20, 29, 35].

Accurate measurements of the PLAX-view biomarkers are crucial for evaluating LV function and identifying cardiac disorders such as hypertrophic cardiomyopathy, myocardial infarction, and heart failure. For instance, increased IVS thickness suggests hypertrophic cardiomyopathy, while the abnormal LVPW measurement is an indicator of myocardial infarction [10]. Additionally, LVID at end-diastole (ED) and end-systole (ES) are fundamental for calculating ejection fraction (EF), a key parameter in diagnosing heart failure [28]. Despite the importance of the PLAX-view in clinical practice, its interpretation is hindered by operator-dependent variability, which compromises the reproducibility of measurements.

2.2. Automated PLAX-view measurement

Automated PLAX-view Echo measurement has emerged as a promising solution to reduce the workload associated with manual measurements and enhance the consistency of clinical assessments. Recent developments in learning-based PLAX-view Echo analysis enable the precise acquisition of key anatomical biomarkers, such as LVID, IVS, and LVPW.

Early techniques, primarily based on convolutional neural networks (CNNs), achieved high performance in detecting key-points of static images for ED and end-systole Echo frames. For instance, EchoNet-LVH [9] employs a DeepLab V3-based static CNN for PLAX-view key-point detection and facilitates automated LV wall quantification. More recent advancements, such as the hierarchical graph neural network (GNN)-based EchoGLAD [30] further demonstrates the potential of learning-based automated Echo measurements.

Despite recent advances, the precise diagnosis of cardiac disorders using deep learning remains challenging. Sophisticated PLAX-view analysis requires evaluating morphological abnormalities across frames within the Echo cycle. However, current learning-based methods are mainly trained on static Echo image datasets, which limits their ability to capture the temporal dynamics of the LV. Therefore, there is a critical need for advanced automated biomarker measurement techniques that ensure enhanced temporal reliability across the Echo cycle.

Video-based Echo analysis enables advanced temporal interpretation of Echo by incorporating inter-frame image understanding. However, video-based analysis remains underexplored in the field of Echo, primarily due to the high cost of key-point annotation for every video frame. To address the limitations, we propose a semi-supervised approach to enhance video-based Echo analysis, utilizing temporal correlations to improve performance.

2.3. Point-tracking

Point-tracking techniques enable the detection and tracking of key-points across video frames. Traditional methods, such as the Kanade-Lucas-Tomasi (KLT) tracker [25, 41], propose to employ optical flow for point trajectory analysis. However, such methods encounter difficulties in complex scenarios involving rapid motion, occlusions, and significant visual changes.

Recent advancements in deep learning have mitigated the limitations of traditional point-tracking methods even under challenging conditions. A prominent contribution to the progress is TAP-Vid [6], which provides diverse datasets for benchmarking point-tracking methods. Deep learning methodologies have also contributed to improving point-tracking accuracy. For instance, RAFT [39], employs dense correlation volumes and iterative updates to achieve enhanced performance. Similarly, CoTracker2 [16], TAPIR and BootsTAP [7, 8], and PIPs/PIPs++ [11, 43] integrate spatial and temporal refinement to achieve enhanced tracking accuracy in complex scenarios.

In Echo, accurate tracking of anatomical deformation is significant for evaluating cardiac function. Recent learning-based point-tracking methods show potential for application in Echo video interpretation, enabling precise key-point trajectory analysis and offering new opportunities for advancing cardiac measurement workflows.

2.4. Semi-supervised video analysis

Semi- and self-supervised video analysis has emerged as a promising paradigm for reducing the cost of dense annotation while simultaneously enhancing model performance [5, 13, 23, 24, 26, 27]. In the domain of Echo, self-supervised strategies propose to exploit the intrinsic periodicity of cardiac motion, thereby learning robust representations from unannotated images through cycle-consistency objectives [5]. As an alternative self-supervision scheme, temporal masking has been employed in Echo [26], pre-training encoders by reconstructing masked frames in temporal sequences, thereby achieving enhanced training of cardiac temporal pattern. Semi-supervised segmentation is predominantly realized through the generation of pseudo labels, by enforcing mean-teacher consistency [38, 42] or by utilizing optical-flow–guided supervision across consecutive frames to encode motion cues [31, 34]. Building on resent advances, we extend such pseudo-label strategy to the task of PLAX-view key-point trajectory estimation. Specifically, we introduce optical-flow–based pseudo key-point labeling to extract supervisory signals from unlabeled Echo videos and to align trajectories without frame-level annotation. We further propose a pixel-wise patch-correlation objective that enforces appearance consistency around each landmark across the cardiac cycle, thereby refining trajectory estimation with higher precision.

3. Method

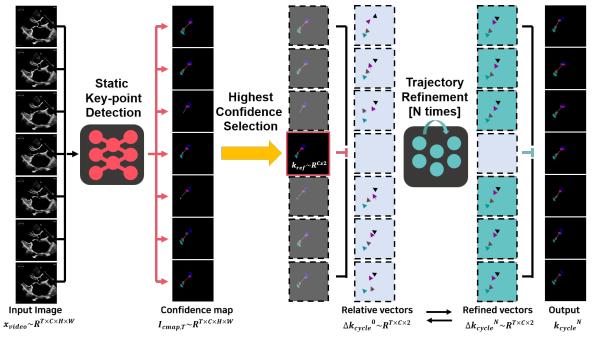

Figure 1. Overview of the proposed Echo key-point trajectory refinement scheme.

Figure 1 presents an overview of the proposed semi-supervised trajectory-refinement scheme. A static key-point detection neural network (SK-net) is employed to generate a confidence map for each frame of the Echo video. The frame exhibiting the highest key-point detection confidence is then selected as the reference frame, with its observed PLAX-view key-points denoted by $k_{\text{ref}} \sim R^{C\times 2}$. C denotes the number of key-points. Then, the relative coordinate vectors for the key-points in each frame are computed with respect to reference key-points, yielding $\Delta k^{0}_{\text{cycle}} \sim \mathbb{R}^{T\times C\times 2}$m where T is the time frames of the Echo. The trajectory refinement network (TR-net) iteratively refines the relative coordinates, resulting in $\Delta k^{N}_{\text{cycle}} \sim \mathbb{R}^{T\times C\times 2}$. By adding the refined coordinates back to $k_{\text{ref}} \sim \mathbb{R}^{C\times 2}$, the refined key-points $k^{N}_{\text{cycle}} \sim \mathbb{R}^{T\times C\times 2}$ is obtained.

3.1. Static key-point detection

To identify key anatomical landmarks for subsequent trajectory analysis, a SK-net is employed. Specifically, the network infers four primary PLAX-view key-points, corresponding to LVID, LVPW, and IVS, from each Echo frame. The network output comprises spatial confidence maps, $I_{\text{cmap}} \sim \mathbb{R}^{C\times H\times W}$, where C is the number of key-points (i.e., 4), and both H and W are set to 224. By analyzing the spatial coordinate showing maximum confidence value for each I_{cmap}, the PLAX key-point locations are extracted as $\mathbf{k} \sim \mathbb{R}^{C\times 2}$. The SK-net architecture is based on HR-net [37], which demonstrates the high computational efficiency and key-point estimation accuracy (discussed in Section 5.1). Training is conducted employing the EchoNet-LVH dataset, which provides static Echo images paired with ground-truth key-points. Each key-point is represented as a heatmap, $I_{\text{GT}} \sim \mathbb{R}^{C\times H\times W}$ generated via Gaussian sampling to emulate human variation.

Due to the sparse nature of the key-point annotations, a modified mean squared error (MSE) loss function is adopted to enhance training stability:

$$\theta_{\text{st}} = \arg\min \mathbb{E}\left[(I_{\text{GT}} - I_{\text{cmap}}) \cdot (1+\gamma)\, I_{\text{GT}}\right], \qquad (1)$$

where $\gamma = 0.01$. The model is optimized using the AdamW optimizer with a learning rate of 1×10^{-4}.

3.2. Trajectory refinement network configuration

Although the SK-net identifies key-points on a frame-by-frame basis, the resulting key-point vectors $\mathbf{k}_{\text{cycle}} \sim \mathbb{R}^{T\times C\times 2}$ do not inherently capture the inter-frame relationships required to characterize the dynamic movement of the LV during systole and diastole. Consequently, tracking LV wall motion and precise interpretation of PLAX-view Echo remains challenging.

To overcome the limitations, we introduce a TR-net designed to leverage inter-frame correlations for enhanced temporal key-point estimation. Specifically, the reference frame, k_{ref}, is selected as the frame that exhibits the highest aggregate I_{cmap} across all key-point channels. TR-net then recursively refines the relative key-point coordinates, Δk_{cycle}, by analyzing the relationships between Echo video frames and corresponding key-point attributes across the cardiac cycle, ensuring enhanced temporal reliability.

Figure 2 provides an overview of the input configuration for the TR-net, which consists of three primary components: (1) relative vectors, (2) semantic features, and (3)

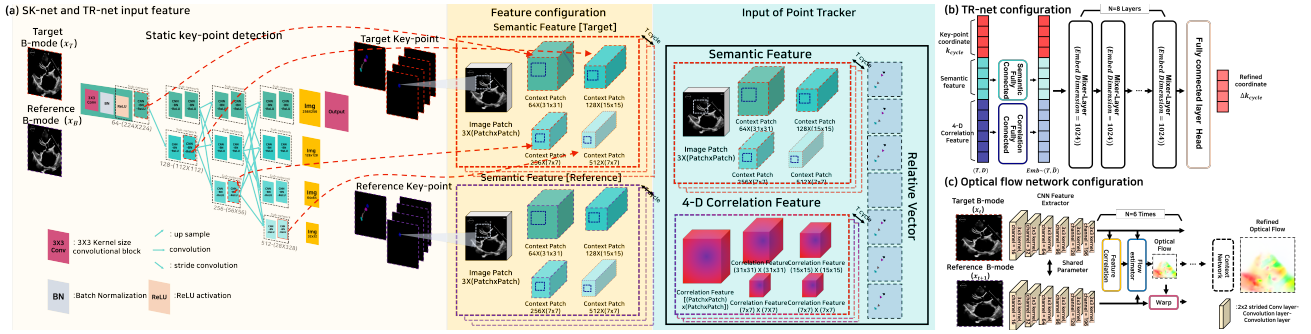

Figure 2. (a) Input configuration of the TR-net. (b) TR-net architecture, (c) Optical flow network configuration.

correlation features. The relative vectors encode the position of each key-point with respect to k_{ref}. The semantic features are subdivided into image patches and context patches. Image patches, widely employed in point-tracking methods, are extracted by cropping $P \times P$ the region around each key-point in every Echo frame, where $P(= 24)$ is the patch size. Context features, introduced as an additional input to the TR-net, are obtained from the intermediate feature representations generated by the SK-net.

As the Echo images are fed forward through the static network (HR-net in this case), perceptual characteristics associated with each key-point are emphasized within the intermediate features. We observed that incorporating the key-point-specific perceptual features into the TR-net significantly improves its ability to track key-points more precisely. In HR-net, as in other convolutional architectures, the image undergoes down-sampling and encoding across multiple resolutions. Accordingly, we extract representative intermediate features $F_{\text{cp}}^{224\times224}$, $F_{\text{cp}}^{112\times112}$, $F_{\text{cp}}^{56\times56}$, $F_{\text{cp}}^{28\times28}$ from the four resolution subnetworks of HR-net (detailed in Figure 2). For each of the intermediate feature maps, a $P \times P$ patch is cropped around the key-point coordinates and employed as a context feature in the trajectory refinement scheme.

In addition to the relative vectors and semantic features, the correlation feature serves as the final input to the TR-net. The correlation feature is specifically designed to capture the correlation between the reference and the target frames, thereby enabling a more accurate estimation of the inter-frame relationship. Following the RAFT framework [39], we configure a 4D correlation feature by computing the dot product of each pixel in the reference frame patch with every pixel in the corresponding target frame patch. Unlike standard implementations that rely solely on image patches, we incorporate context patches to embed the correlation information with perceptual cues derived from intermediate representations of SK-net, enhancing the overall performance of TR-net.

The network operates recursively, iteratively refining $\Delta k_{\text{cycle}}^{i}$. At each iteration i, the key-point coordinates for each frame are updated according to

$$k_{\text{cycle}}^{i} = k_{\text{ref}} + \Delta k_{\text{cycle}}^{i}. \tag{2}$$

Using the updated coordinates, the corresponding semantic features, image and context patches, and the correlation feature are recalculated for the subsequent iteration. Through the iterative process, key-point localization is progressively improved, thereby facilitating robust tracking of LV wall movement throughout the cardiac cycle.

3.3. Semi-supervised trajectory learning scheme

We introduce a semi-supervised trajectory learning scheme, which enables the TR-net to be trained using unannotated Echo videos.

To enable semi-supervised training of the TR-net, we introduce two loss functions: (i) pseudo optical flow loss and (ii) weighted patch correlation loss. During each training iteration, the network dynamically balances the relative contributions of these loss functions, effectively utilizing their complementary properties.

Pseudo optical flow loss utilizes a pretrained optical flow estimation model, adapted for Echo video, to generate pseudo key-point labels $\bar{k}_{\text{cycle}}$. At the early stages of each iteration, greater weight is given to the pseudo optical flow loss, thereby aligning the overall trajectory with the pseudo labels. The alignment step serves as a coarse yet stable initialization for subsequent refinements. In parallel, the weighted patch correlation loss measures the pixel-wise similarity of image patches centered on key-points across frames. By leveraging key-point patch features, the loss enables a more precise estimation of key-point trajectories. As iteration progresses, the weight of the patch correlation loss gradually increases, allowing the TR-net to achieve progressively refined trajectory adjustments with each iteration.

Pseudo optical flow loss. Figure 3 illustrates the configuration of the pseudo optical flow loss, which is introduced to align the TR-net predictions with pseudo key-point labels. To generate the pseudo labels, we employed an optical flow neural network. Figure 2. (c) shows configuration of the network architecture, which is based on PWCNet

[36]. The target and reference B-mode frames are processed through a shared-weight feature extractor. Then, a series of feature correlation and flow estimators are applied to progressively refine the optical flow. The network is trained using the unsupervised framework proposed in ARFlow [22]. EchoNet-LVH video data are used to train the optical flow model, ensuring that the network is adapted to the characteristics of Echo images.

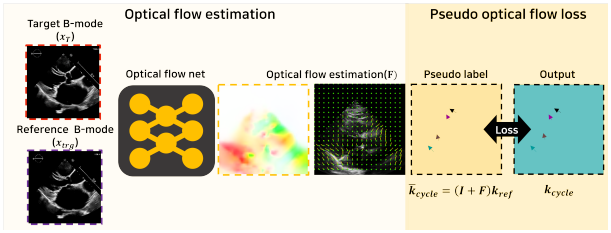

Figure 3. Illustration of the pseudo optical flow loss framework.

The SK-net provides a reference Echo frame as introduced in section 3.1. The pretrained optical flow network measures the optical flow, $\mathbf{F}$, between the reference frame and target frame in the Echo cycle. The optical flow is then applied to the reference key-points, $\mathbf{k}_{\text{ref}}$, generating the pseudo key-point labels for the target frame $\bar{\mathbf{k}}_{\text{cycle}} = (\mathbf{I} + \mathbf{F})\mathbf{k}_{\text{ref}}$, where $\mathbf{I}$ denotes the identity matrix.

The network is trained to minimize squared loss between $\mathbf{k}^i_{\text{cycle}}$ and $\bar{\mathbf{k}}_{\text{cycle}}$. The loss function is denoted as follows:

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{N} \sum \left(\bar{\mathbf{k}}_{\text{cycle}} - \mathbf{k}^i_{\text{cycle}}\right)^2. \tag{3}$$

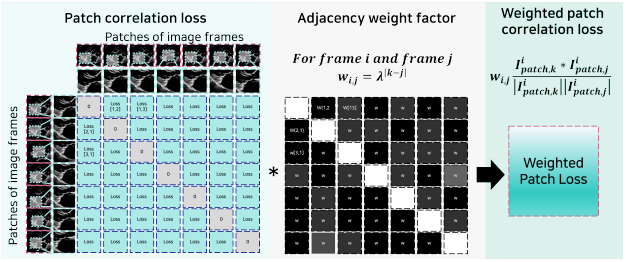

Figure 4. Configuration of weighted patch loss.

Weighted patch correlation loss. Although the pseudo optical flow loss provides a coarse alignment based on optical flow-derived pseudo labels, it alone is insufficient to achieve precise trajectory refinement. To address the limitations, we propose a weighted patch correlation loss, illustrated in Figure 4, which directly measures the similarity of the image patches surrounding each key-point across the Echo cycle.

Specifically, for each frame in the cycle, we extract a patch centered on the key-point prediction $\mathbf{k}^i_{\text{cycle}}$. We then compute the cosine similarity between the patches across all pairs of frames, assigning higher weights to pairs of frames that are closer in time. The proposed weighting scheme exploits the observation that adjacent frames tend to preserve local similarities more robustly, thereby effective in local trajectory refinement. The weighted patch correlation loss is defined as:

$$\mathcal{L}_{\text{patch}} = \lambda^{|k-j|} \frac{I^i_{\text{patch},k} \cdot I^i_{\text{patch},j}}{\|I^i_{\text{patch},k}\| \|I^i_{\text{patch},j}\|}, \tag{4}$$

where $\lambda = 0.9$ adjusts the penalty according to the temporal distance $|k - j|$. Here, $I^i_{\text{patch},k}$ and $I^i_{\text{patch},j}$ denote the patches extracted around the key-point in frames k and j, respectively.

Refinement over iterations. The TR-net progressively updates the predicted key-points over multiple iterations, transitioning from coarse alignment to a more precise trajectory estimation. To facilitate the refinement, we combine the pseudo optical flow loss $\mathcal{L}_{\text{pseudo}}$ and the weighted patch correlation loss $\mathcal{L}_{\text{patch}}$ with iteration-dependent weighting factors. The overall loss function at each iteration is defined as:

$$\mathcal{L} = \alpha^{|M-i_t|} \left(\mathcal{L}_{\text{pseudo}} + \beta^{|M-i_t|} \mathcal{L}_{\text{patch}}\right) \tag{5}$$

where hyper-parameter $\alpha = 0.8$ and $\beta = 0.8$. $M = 5$ denotes the total number of iterations, and i_t is the index of current iteration.

In the initial iteration phase, $\mathcal{L}_{\text{pseudo}}$ is weighted heavily, enabling the network to converge efficiently on a coarse global trajectory derived from the optical flow–based pseudo labels. As iteration progresses, the weight of $\mathcal{L}_{\text{patch}}$ increases, refining local key-point alignment via the patch similarity measurement.

Implementation details. MLP-Mixer [40] is employed as the backbone architecture of TR-net. The configuration of TR-net is presented in Figure 2. Semantic and correlation features are embedded through fully connected layers and concatenated with k^i_{cycle}. The concatenated representation is subsequently processed through a series of mixer layers to generate $\Delta k^i_{\text{cycle}}$. AdamW optimizer with a learning rate of le-4 is applied for the optimization. The open-access EchoNet-LVH Echo videos are employed for the semi-supervised training of the TR-net.

4. Dataset

EchoNet-LVH dataset. EchoNet-LVH is an open-access Echo video dataset specifically designed for the evaluation of left ventricular hypertrophy (LVH). The dataset comprises 12,000 echocardiogram videos. For each video, up to two static frames representing end-diastole and end-systole are annotated with four key-points to offer precise measurement of LVID, IVS, and LVPW. EchoNet-LVH is developed

Table 1. Quantitative assessments of TR-net and baseline scheemes. Metrics include the $< \delta_{avg}$, MAE and AVE for Septal Endocardial Point (SEnP), Septal Epicardial Point (SEpP), Posterior wall Endocardial Point(PEnP), and Posterior wall Epicardial Point (PEpP)

	Name	Method		MAE			MAE (ED/ES)			$\sigma_{temporal}$			AVE			
			$< \delta_{avg}$	LVPWd	LVIDs	LVSd	LVPWd	LVIDs	LVSd	LVPWd	LVIDs	LVSd	SEnP	SepP	PEnP	PEpP
Proposed	**SK-net**	**Static Supervised**	30.3%	5.03	*9.18*	6.15	4.71	*8.73*	5.25	1.41	1.77	1.46	3.33	3.46	2.23	1.70
	TR-net	**Point tracking**	**32.8%**	*4.83*	**8.93**	**3.35**	*4.69*	**8.54**	**3.50**	*1.12*	**1.72**	**0.98**	**1.35**	*1.62*	**1.13**	**0.91**
Conventional	**KLT Tracker**	**Point tracking**	25.9%	7.86	11.40	4.44	7.80	11.59	5.11	4.53	2.99	2.58	1.54	1.97	1.29	1.01
	Speckle Tracking	**Point tracking**	26.7%	6.09	10.85	3.84	5.68	10.46	4.01	1.64	1.98	1.23	1.46	*1.60*	*1.19*	0.94
Echo-based TM	**Echonet-LVH**	**Static Supervised**	25.0%	7.93	12.64	*3.46*	5.50	12.69	4.29	1.37	2.15	*1.13*	3.73	3.79	2.89	2.13
	EchoGLAD	**GNN**	26.2%	13.31	10.75	10.15	19.58	14.33	12.20	2.51	2.50	1.62	4.79	4.78	4.05	2.53
	EchoTracker	**Point tracking**	*31.3%*	**4.66**	9.81	3.67	**4.51**	9.87	4.00	**1.11**	1.85	1.19	*1.42*	1.65	1.33	1.02
CV-based TM	**CoTracker3 offline**	**Point tracking**	29.5%	6.69	10.48	4.52	6.34	10.44	5.00	5.00	4.03	2.50	1.63	2.11	1.57	1.80
	CoTracker3 online	**Point tracking**	26.4%	6.97	12.93	4.51	6.71	13.48	5.39	4.85	7.48	2.38	2.24	2.82	2.30	2.00
	CoTracker2	**Point tracking**	20.6%	8.71	18.56	6.10	9.25	21.02	7.29	5.85	6.13	2.71	3.69	3.92	3.28	3.25
	BootsTAPIR	**Point tracking**	30.5%	5.12	9.58	3.51	4.94	11.20	*3.88*	3.50	*1.75*	1.17	1.46	**1.55**	1.29	*0.93*
	TAPIR	**Point tracking**	18.6%	12.29	24.56	7.90	11.90	24.99	9.37	3.62	5.51	2.43	4.97	4.19	2.77	2.76

by Stanford University researchers, who adhered to strict ethical standards.

PLAX video dataset. For precise evaluation of Echo video key-point trajectory schemes, we configured a PLAX-view Echo video dataset. The proposed PLAX video dataset includes annotations for four key-points in every frame of the PLAX-view Echo video. The Echo videos are acquired using a diverse range of ultrasound machines, including the Vivid E9 and Vivid IQ series from GE Healthcare (US), the EPIQ 7C and EPIQ CVx systems from Philips Medical Systems (NL), and the ACUSON SC2000 from Siemens (DE).

Table 2. Distribution of cardiac diseases

Disease category	# subject	Age (mean ± SD)	Male	Female
Normal	75	64.3 ± 13.8	35 (46.7%)	40 (53.3%)
Angina	62	67.4 ± 10.9	37 (59.7%)	25 (40.3%)
Atrial fibrillation	32	64.9 ± 16.4	16 (50.0%)	16 (50.0%)
Cardiomyopathy	16	67.6 ± 10.1	9 (56.2%)	7 (43.8%)
Heart failure	12	59.9 ± 15.7	10 (83.3%)	2 (16.7%)
Myocardial infarction	12	72.9 ± 14.2	8 (66.7%)	4 (33.3%)
Valvular disease	7	73.7 ± 8.3	1 (14.3%)	6 (85.7%)
Others	82	61.7 ± 14.6	40 (48.8%)	42 (51.2%)
Total	298	64.9 ± 14.0	156 (52.3%)	142 (47.7%)

For each subject, comprehensive diagnostic procedures, including MRI, CT, and patient history reviews, are conducted and patient's cardiac disease is confirmed by medical experts. Detailed subject configurations, including age and gender distribution across different disease categories, are introduced in Table 2. Annotations are performed by six medical experts with extensive experience in Echo.

The annotations are verified and refined using full video sequences to ensure accurate tracking and coherence of the labeled points throughout the cardiac cycle. For further refinement of the dataset, the labeled data is reviewed by a clinician, ensuring the reliability of the dataset. The dataset comprises 298 US videos. Each video includes more than one full cardiac cycle. The video consists of a temporal sequence with more than 24 image frames. The image resolution of the Echo video is 224 × 224 pixels.

5. Experiments

Evaluation metrics. For quantitative assessment of the proposed TR-net and the comparative baselines, we employed the average distance within the threshold, δ_{avg}, as proposed in TAP-Vid [6]. We report the mean average error (MAE) of LVID, IVS, and LVPW of entire cardiac cycle and ED and ESD key cardiac frames, to evaluate the accuracy of the biomarker measurements. Temporal stability is assessed by measuring standard deviation $\sigma_{temporal}(\cdot)$ between the ground-truth key-point LVID, IVS, and LVPW and the reconstructed ones over time. In addition, the average vector error (AVE), calculated as the average magnitude of the difference vector between the predicted and ground truth points over time, is investigated.

5.1. Comparative study

In this section, we present a comparative study evaluating the performance of the proposed TR-net for key-point estimation in Echo videos. The proposed method is benchmarked against multiple state-of-the-art (SoTA) baselines. The comparative baselines include conventional speckle tracking scheme [14], KLT tracker [4], the the EchoNet-LVH network [9], and EchoGLAD [30]. In addition, we compare the TR-net with EchoTracker [3], which applies point-tracking techniques to Echo.

We also evaluate advanced point-tracking networks from the computer vision domain, including TAPIR [7], BootsTAP [8], Cotracker2 [16], and Cotracker3 [15], which are known for robust performance in tracking tasks.

Quantitative assessment. Quantitative results, summarized in Table 1, indicate that the proposed model outperforms the baseline methods in overall. In the static inference experiment, the proposed SK-net achieves a higher δ_{avg} and lower MAE than -EchoNet-LVH. The result demonstrates the efficacy of the SK-net architecture introduced in Section 3.1. Leveraging TR-net for the key-point trajectory refinement significantly improves key-point estimation accuracy, achieving a 16% reduction in MAE compared to SK-net.

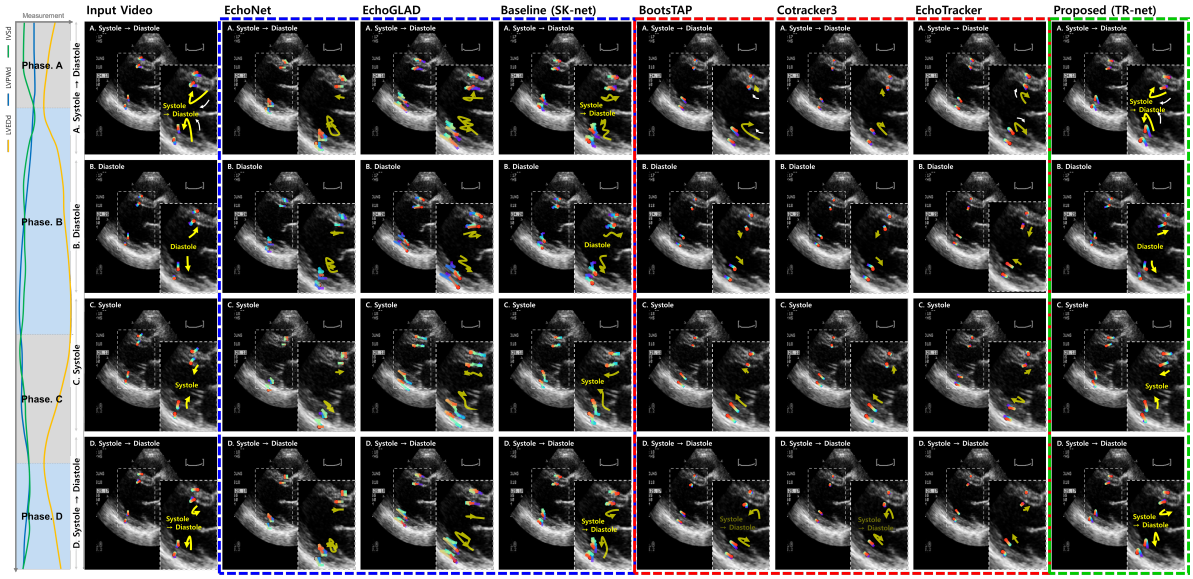

Figure 5. Qualitative assessments. The key-point trajectories are analyzed across four cardiac phases (Phase A, B, C, and D).

The proposed TR-net demonstrates enhanced robustness compared to the GNN-based EchoGLAD showing 6.6% higher δ_{avg}. When evaluated against EchoTracker, the proposed semi-supervised trajectory estimation approach exhibites improved performance in δ_{avg} and AVE metrics, effectively interpreting the dynamic movements of Echo.

Among computer vision-based tracking methods, BootsTAP achieves the best performance, demonstrating 30.5 δ_{avg}. However, the proposed TR-net outperforms BootsTAP in both δ_{avg} and MAE, emphasizing the suitability of the semi-supervised learning paradigm for echocardiographic key-point tracking. In addition, the proposed TR-net demonstrates a improvement in temporal stability compared with the baseline static SK-net. In addition, TR-net shows superior performance relative to other comparative networks.

Qualitative assessment. Figure 5 provides a qualitative assessment of the proposed TR-net and comparative baselines. Static models, including EchoNet-LVH, EchoGLAD, and the baseline SK-net, accurately identify key-points in individual frames but lack temporal consistency, resulting in unstable and jittery trajectories. Temporal models such as BootsTAP, Cotracker3, and EchoTracker demonstrate improved temporal consistency leveraging inter-frame correlations. However, the networks struggled to capture the dynamic movements and rapid transition of the heart during systole and diastole. In contrast, the proposed model, despite its semi-supervised nature, effectively interprets semantic information and inter-frame correlations to accurately detect cardiac wall movements. The proposed model reliably identifies clinically significant landmarks, including LVID, IVS, and LVPW. The results validate the proposed TR-net as a robust and clinically applicable solution for echocardiographic key-point tracking.

5.2. Ablation study

Table 3. Ablation study on the loss configuration

	L_{pseudo}	L_{patch}	$< \delta_{\text{avg}} \uparrow$	MAE		
				LVPWd	LVIDs	LVSd
Proposed			30.3%	5.03	9.18	6.15
	o		*31.3%*	4.99	9.04	3.41
	o	o	**32.8%**	**4.83**	**8.93**	**3.35**

Loss configuration. An ablation study is conducted to evaluate the contributions of the proposed $\mathcal{L}_{\text{pseudo}}$ and $\mathcal{L}_{\text{patch}}$ to the model performance. The quantitative results in Table 3 present the efficacy of loss configuration. Training the model exclusively with $\mathcal{L}_{\text{patch}}$ makes the optimization challenging. The combined use of $\mathcal{L}_{\text{pseudo}}$ and $\mathcal{L}_{\text{patch}}$ presents the optimal results, achieving the highest δ_{avg} (32.8%) and the lowest MAE values for LVPWd (4.83), LVIDs (8.93), and LVSd (3.35). The $\mathcal{L}_{\text{pseudo}}$ provides global trajectory alignment for stabilizing the trajectory refinement, while $\mathcal{L}_{\text{patch}}$ enables the network to refine the detailed key-point trajectory effectively.

Figure 6 demonstrates the accuracy of the proposed scheme with respect to iteration progression and hyperparameter variations. In terms of performance across iterations, the baseline model, with the ablation of $\mathcal{L}_{\text{patch}}$, rapidly converged at early iterations, resulting in limited further performance improvement. However, the proposed scheme, which utilizes $\mathcal{L}_{\text{patch}}$ to more precisely analyze the inter-frame relationships between key-points, showed consistent

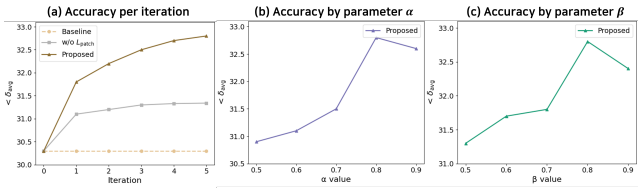

Figure 6. Quantitative assessment. (a) Accuracy plot per iteration. Accuracy plot for varying parameter α (b) and β (c)

improvements in accuracy as iterations progressed. Regarding the hyperparameter analysis, optimal accuracy was empirically confirmed at $\alpha = 0.8$ and $\beta = 0.8$.

Table 4. Ablation study of the input configuration.

	Semantic input	$< \delta_{\text{avg}} \uparrow$	MAE LVPWd	MAE LVIDs	MAE LVSd
Proposed		31.3%	4.87	8.97	3.37
	o	**32.8%**	**4.83**	**8.93**	**3.35**

Input configuration. The influence of semantic features on the input configuration is assessed through the ablation experiment. Table 4 presents the quantitative evaluation of the ablation study. Incorporating semantic features enhances the perceptual understanding of individual key-points, leading to improved localization accuracy. The results demonstrate that integrating semantic features improves the TR-net key-point detection precision by 2.5% in the δ_{avg} metric.

Table 5. Quantitative assessment of the proposed confidence-based frame configuration.

	Frame	$< \delta_{\text{avg}} \uparrow$	MAE LVPWd	MAE LVIDs	MAE LVSd
Proposed	**First**	32.4%	5.22	9.71	*3.52*
	Random	*32.7%*	*5.08*	*9.65*	3.78
	ours	**32.8%**	**4.83**	**8.93**	**3.35**

Confidence-based frame configuration. The proposed model incorporates a confidence-based frame selection scheme, where the frame with the highest key-point confidence is designated as the reference frame for key-point tracking. To assess the effectiveness, two alternative configurations are made: selecting the first frame as the reference frame, a common practice in general computer vision tasks, and selecting a frame at random. The results indicate that the confidence-based frame configuration provides an optimal initial condition for point tracking, leading to enhanced overall trajectory estimation performance.

Table 6. Comparison of methods on three cardiac diseases

	Heart failure			Cardiomyopathy			Myocardial infarction		
	AUC	Spe	Sen	AUC	Spe	Sen	AUC	Spe	Sen
TR-net	0.85	0.77	0.70	0.83	0.68	0.80	0.82	0.80	0.73
SK-net	0.82	0.70	0.69	0.73	0.60	0.60	0.75	0.67	0.70
Kumar, et al	0.75	0.73	0.63	0.70	0.65	0.70	0.62	0.60	0.60
Speckle tracking	0.78	0.67	0.63	0.66	0.59	0.63	0.75	0.67	0.67
EchoNet-LVH	0.66	0.61	0.43	0.76	0.40	0.60	0.65	0.60	0.67
EchoTracker	0.70	0.73	0.57	0.71	0.40	0.60	0.73	0.67	0.73
CoTracker3 offline	0.73	0.65	0.63	0.62	0.57	0.53	0.73	0.60	0.67

6. Cardiac disease classification

To assess the clinical effectiveness of the proposed framework, we conducted an evaluation on representative cardiac conditions present within the dataset, namely heart failure, cardiomyopathy, and myocardial infarction. For each method under comparison, the lengths of the IVS, left LVID, and LVPW are measured across the entire cardiac cycle, beginning at ED. The scheme of Kumar et al. [17] is evaluated by implementing the scale-invariant feature extraction method described in the study. These quantitative features are subsequently used as inputs to a support vector machine classifier to distinguish between normal and pathological subjects, with performance assessed via 5-fold cross-validation. Evaluation metrics include the AUROC, specificity (Spe), and sensitivity (Sen), as summarized in Table 6. By providing refined key-point trajectories, TR-net demonstrates superior performance in downstream cardiac disease classification, achieving higher AUROC not only compared with conventional pre-deep learning approaches but also relative to recent learning-based automated Echo analysis approaches.

Limitations. While the TR-net improves spatiotemporal consistency, several limitations persist. Performance may deteriorate in studies with suboptimal acoustic windows that partially occlude the myocardial wall and weaken local evidence. Robust trajectory estimation in the presence of occlusions is a promising subject in further investigation.

7. Conclusion

In this paper, we introduce a semi-supervised framework for key-point trajectory analysis in Echo videos, addressing the critical need for temporally consistent tracking of LV key-points. By incorporating the proposed semi-supervised learning strategy, TR-net achieves enhanced performance in spatial precision and temporal stability. The quantitative and qualitative evaluations validate the efficacy of the method, emphasizing the potential for clinical deployment in Echo workflows.

8. Ethics statement

All patient data used in this study are obtained under the IRB approval of Seoul National University Bundang Hospital (IRB Number: B-2305-828-104).

References

[1] Fabio Angeli, Paolo Verdecchia, Enrica Angeli, Fabrizio Poeta, Mariagrazia Sardone, Maurizio Bentivoglio, Lucio Prosciutti, Maurizio Cocchieri, Liliana Zollino, Gianni Bellomo, et al. Day-to-day variability of electrocardiographic diagnosis of left ventricular hypertrophy in hypertensive patients. influence of electrode placement. *Journal of Cardiovascular Medicine*, 7(11):812–816, 2006. 1

[2] Toshihiko Asanuma, Ayumi Uranishi, Kasumi Masuda, Fuminobu Ishikura, Shintaro Beppu, and Satoshi Nakatani. Assessment of myocardial ischemic memory using persistence of post-systolic thickening after recovery from ischemia. *JACC: Cardiovascular Imaging*, 2(11):1253–1261, 2009. 1

[3] Md Abulkalam Azad, Artem Chernyshov, John Nyberg, Ingrid Tveten, Lasse Lovstakken, Håvard Dalen, Bjørnar Grenne, and Andreas Østvik. Echotracker: Advancing myocardial point tracking in echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 645–655. Springer, 2024. 6

[4] Jean-Yves Bouguet et al. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation*, 5(1-10):4, 2001. 6

[5] W Dai, X Li, X Ding, and KT Cheng. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos (2022). 3

[6] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2, 6

[7] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 2, 6

[8] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pages 3257–3274, 2024. 2, 6

[9] Grant Duffy, Paul P Cheng, Neal Yuan, Bryan He, Alan C Kwan, Matthew J Shun-Shin, Kevin M Alexander, Joseph Ebinger, Matthew P Lungren, Florian Rader, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA cardiology*, 7(4):386–395, 2022. 1, 2, 6

[10] Perry M Elliott, Aris Anastasakis, Michael A Borger, Martin Borggrefe, Franco Cecchi, Phillippe Charron, Albert Alain Hagege, Antoine Lafont, Giuseppe Limongelli, Heiko Mahrholdt, et al. 2014 esc guidelines on diagnosis and management of hypertrophic cardiomyopathy. *Polish Heart Journal (Kardiologia Polska)*, 72(11):1054–1126, 2014. 2

[11] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 2

[12] James P Howard, Catherine C Stowell, Graham D Cole, Kajaluxy Ananthan, Camelia D Demetrescu, Keith Pearce, Ronak Rajani, Jobanpreet Sehmi, Kavitha Vimalesvaran, G Sunthar Kanaganayagam, et al. Automated left ventricular dimension assessment using artificial intelligence developed and validated by a uk-wide collaborative. *Circulation: Cardiovascular Imaging*, 14(5):e011951, 2021. 1

[13] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 690–706, 2018. 3

[14] Philippe Joos, Jonathan Porée, Hervé Liebgott, Didier Vray, Mathilde Baudet, Julia Faurie, François Tournoux, Guy Cloutier, Barbara Nicolas, and Damien Garcia. High-frame-rate speckle-tracking echocardiography. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 65(5): 720–728, 2018. 6

[15] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 6

[16] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2025. 2, 6

[17] Ritwik Kumar, Fei Wang, David Beymer, and Tanveer Syeda-Mahmood. Cardiac disease detection from echocardiogram using edge filtered scale-invariant motion features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 162–169. IEEE, 2010. 8

[18] Patrizio Lancellotti, Susanna Price, Thor Edvardsen, Bernard Cosyns, Aleksandar N Neskovic, Raluca Dulgheru, Frank A Flachskampf, Christian Hassager, Agnes Pasquet, Luna Gargani, et al. The use of echocardiography in acute cardiovascular care: recommendations of the european association of cardiovascular imaging and the acute cardiovascular care association. *European Heart Journal-Cardiovascular Imaging*, 16(2):119–146, 2015. 1

[19] Roberto M Lang, Michelle Bierig, Richard B Devereux, Frank A Flachskampf, Elyse Foster, Patricia A Pellikka, Michael H Picard, Mary J Roman, James Seward, Jack S Shanewise, et al. Recommendations for chamber quantification: a report from the american society of echocardiography's guidelines and standards committee and the chamber quantification writing group, developed in conjunction with the european association of echocardiography, a branch of the european society of cardiology. *Journal of the American society of echocardiography*, 18(12):1440–1463, 2005. 2

[20] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update

from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015. 1, 2

[21] Emily S Lau, Paolo Di Achille, Kavya Kopparapu, Carl T Andrews, Pulkit Singh, Christopher Reeder, Mostafa Al-Alusi, Shaan Khurshid, Julian S Haimovich, Patrick T Ellinor, et al. Deep learning–enabled assessment of left heart structure and function predicts cardiovascular outcomes. *Journal of the American College of Cardiology*, 82(20): 1936–1948, 2023. 1

[22] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6489–6498, 2020. 5

[23] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6489–6498, 2020. 3

[24] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4571–4580, 2019. 3

[25] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2

[26] Fadillah Maani, Asim Ukaye, Nada Saadi, Numan Saeed, and Mohammad Yaqub. Simlvseg: simplifying left ventricular segmentation in 2-d+ time echocardiograms with self-and weakly supervised learning. *Ultrasound in Medicine & Biology*, 50(12):1945–1954, 2024. 3

[27] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[28] Authors/Task Force Members:, Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, et al. 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc). with the special contribution of the heart failure association (hfa) of the esc. *European journal of heart failure*, 24 (1):4–131, 2022. 2

[29] Carol Mitchell, Peter S Rahko, Lori A Blauwet, Barry Canaday, Joshua A Finstuen, Michael C Foster, Kenneth Horton, Kofo O Ogunyankin, Richard A Palma, and Eric J Velazquez. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019. 2

[30] Masoud Mokhtari, Mobina Mahdavi, Hooman Vaseli, Christina Luong, Purang Abolmaesumi, Teresa SM Tsang, and Renjie Liao. Echoglad: Hierarchical graph neural networks for left ventricle landmark detection on echocardiograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 227–237. Springer, 2023. 1, 2, 6

[31] Siva Karthik Mustikovela, Michael Ying Yang, and Carsten Rother. Can ground truth label propagation from video help semantic segmentation? In *European conference on computer vision*, pages 804–820. Springer, 2016. 3

[32] Patricia A Pellikka, Adelaide Arruda-Olson, Farooq A Chaudhry, Ming Hui Chen, Jane E Marshall, Thomas R Porter, and Stephen G Sawada. Guidelines for performance, interpretation, and application of stress echocardiography in ischemic heart disease: from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 33(1):1–41, 2020. 1

[33] Dermot Phelan, Brett W Sperry, Paaladinesh Thavendiranathan, Patrick Collier, Zoran B Popović, Harry M Lever, Nicholas G Smedira, and Milind Y Desai. Comparison of ventricular septal measurements in hypertrophic cardiomyopathy patients who underwent surgical myectomy using multimodality imaging and implications for diagnosis and management. *The American journal of cardiology*, 119(10): 1656–1662, 2017. 1

[34] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint learning of motion estimation and segmentation for cardiac mr image sequences. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 472–480. Springer, 2018. 3

[35] Nelson B Schiller, Pravin M Shah, Michael Crawford, Anthony DeMaria, Richard Devereux, Harvey Feigenbaum, Howard Gutgesell, Nathaniel Reichek, David Sahn, Ingela Schnittger, et al. Recommendations for quantitation of the left ventricle by two-dimensional echocardiography. *Journal of the American Society of Echocardiography*, 2(5):358–367, 1989. 2

[36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 5

[37] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV*

2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020. 2, 4

[40] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 5

[41] Carlo Tomasi and Takeo Kanade. Detection and tracking of point. *Int J Comput Vis*, 9(137-154):3, 1991. 2

[42] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022. 3

[43] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 2

[44] S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019. 1