Motion estimation by deep learning in 2D echocardiography: synthetic dataset and validation

Ewan Evain, Yunyun Sun, Khuram Faraz, Damien Garcia, Eric Saloux, Bernhard L Gerber, Mathieu de Craene, Olivier Bernard

Motion estimation by deep learning in 2D echocardiography: synthetic dataset and validation

Ewan Evain, Yunyun Sun, Khuram Faraz, Damien Garcia, Eric Saloux, Bernhard L. Gerber, Mathieu De Craene, and Olivier Bernard

***Abstract*— Motion estimation in echocardiography plays an important role in the characterization of cardiac function, allowing the computation of myocardial deformation indices. However, there exist limitations in clinical practice, particularly with regard to the accuracy and robustness of measurements extracted from images. We therefore propose a novel deep learning solution for motion estimation in echocardiography. Our network corresponds to a modified version of PWC-Net which achieves high performance on ultrasound sequences. In parallel, we designed a novel simulation pipeline allowing the generation of a large amount of realistic B-mode sequences. These synthetic data, together with strategies during training and inference, were used to improve the performance of our deep learning solution, which achieved an average endpoint error of** 0.07 ± 0.06 **mm per frame and** 1.20 ± 0.67 **mm between ED and ES on our simulated dataset. The performance of our method was further investigated on 30 patients from a publicly available clinical dataset acquired from a GE system. The method showed promise by achieving a mean absolute error of the global longitudinal strain of** $2.5 \pm 2.1\%$ **and a correlation of** 0.77 **compared to GLS derived from manual segmentation, much better than one of the most efficient methods in the state-of-the-art (namely the FFT-Xcorr block-matching method). We finally evaluated our method on an auxiliary dataset including 30 patients from another center and acquired with a different system. Comparable results were achieved, illustrating the ability of our method to maintain high performance regardless of the echocardiographic data processed.**

***Index Terms*— Deep learning, Echocardiography, Motion Estimation, Ultrasound Imaging**

I. Introduction

ULTRASOUND imaging is a widely used imaging modality in cardiology because it is inexpensive, fast and non-invasive. Echocardiography enables the extraction of clinical indices relevant to study the cardiac function and anatomy such as volumes and myocardial deformation [1]. Deformation indices are usually estimated by conventional motion estimation techniques that suffer from difficulties inherent in ultrasound images, such as artifacts (shadow, reverberation), lack of information or speckle decorrelation. The latter corresponds to the fact that the speckle pattern which is tracked from B-mode sequences can change over time. This phenomenon depends on the type of movement of the tissues (rotation being one of the worst) and is all the more true as the movements are fast [2]. This results in a lack of accuracy and reproducibility in current embedded solutions. Therefore, improvements in motion estimation are crucial in ultrasound imaging to obtain reproducible indices. One index that has attracted considerable attention is the global longitudinal strain (GLS). GLS is defined as the percentage of myocardial longitudinal shortening between the end-diastolic and end-systolic instants [3]. It is a global value that proved to be robust enough to be part of the recommendations during clinical exams [4]. GLS is computed from B-mode images acquired in any standard apical view and by tracking a myocardial contour using the conventional block-matching [5] or optical flow techniques [6]. Tissue Doppler techniques can also be used to estimate GLS without the use of speckle tracking.

Deep learning (DL) approaches have recently outperformed standard tracking methods on natural images. In particular, the benchmark on the Sintel dataset [1] shows that the top-ranked algorithms are all based on DL approaches and that the first non-DL method (FlowFields [7]) is currently ranked above 100. We thus hypothesized that DL can significantly improve tracking accuracy and robustness over traditional methods in ultrasound imaging. Instead of relying only on the intensity or the phase information in the image to evaluate the motion, DL networks can learn to estimate complex tissue motion with the associated speckle decorrelation. Moreover, the addition of typical ultrasound artifacts during training should provide greater robustness of motion estimation and better adaptation to ultrasound images.

FlowNet [8] was the first neural network trained end-to-end to predict the optical flow from a pair of images. FlowNet consisted of two separate networks: FlowNet-S based only on U-Net [9] and FlowNet-C, which introduced the notion of a cost volume block merging layers from the contraction part

This work was supported by the ANRT (Agence Nationale de la Recherche et de la Technologie) through the CIFRE program.

E. Evain, Y. Sun, K. Faraz, D. Garcia and O. Bernard are with the University of Lyon, CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, University of Lyon 1, Villeurbanne, France.

E. Evain and M. De Craene are with Philips Research Paris (Medisys), Suresnes, France (e-mail: ewan.evain@philips.com)

E. Saloux is with Normandie University, UNICAEN, CHU de Caen Normandie, Department of Cardiology, EA4650 SEILIRM, Caen, France

B.L. Gerber is with Cliniques Universitaires Saint-Luc UCL, Brussels, Belgium

[1] http://sintel.is.tue.mpg.de/results

of the network. In FlowNet2 [10], a new network, FlowNet-SD, close to the FlowNet-S structure was introduced to better manage small displacements. By stacking these different networks with intermediate warpings, FlowNet2 outperformed the state-of-the-art methods but used 160M parameters. SpyNet [11] reduced the number of parameters to 1.2M by using a coarse to fine pyramidal network with image warping. Performances were on par with those of FlowNet but under those of FlowNet2. Finally, PWC-Net [12] obtained better results by combining the pyramid structure of SpyNet, the cost volume as in FlowNet-C and a warping step realized on the feature maps.

The networks mentioned above have been applied to cardiac imaging, with a main focus on MRI [13]–[16]. In parallel, a pilot study has recently shown the adaptability of FlowNet-based networks to the characteristics of ultrasound images for motion estimation [2]. Most methods in ultrasound have been applied to elastography [17], [18], among which some are based on extensions of key architectures such as PWC-Net [19], [20]. Some studies have also been conducted to estimate myocardial motion in echocardiographic imaging. In [21], an unsupervised approach based on the U-Net architecture was used to estimate the canine myocardial motion from a short-axis view. Evaluated on the same data, another network with an architecture derived from FlowNet-C was developed and trained in a semi-supervised way [22]. Short axis view essentially provided information on radial and circumferential strain. To track longitudinal motion, a pipeline was implemented to automate the GLS computation using view classification, segmentation, motion estimation and Kalman filters on apical four chambers views [23]. The motion estimation part was based on FlowNet2 with the original network weights learned from natural synthetic images. Recently, another pipeline with a modified version of PWC-Net named EchoPWC-Net was introduced [24]. To adapt this network to ultrasound images, the authors removed the feature maps warping, propagated the first feature maps and added finer resolutions to the loss. This network was trained on a realistic simulated ultrasound dataset [25] in a supervised way and evaluated on the same *in-silico* dataset and on 30 *in-vivo* patients. Despite all the architectural modifications, the clinical measurements obtained on the real data were only slightly better than those obtained by a state-of-the-art method. Based on these results, the authors highlighted the importance of simulated data and pointed out the lack in quantity and diversity of training data currently available.

A. Main contributions

This paper makes contributions regarding the PWC-Net architecture, synthetic training data for capturing motion in ultrasound, a thorough investigation of different temporal strategies for improving results, and the first study on the generalization of this type of network in echocardiography:

- To overcome the problem of limited synthetic data in number and diversity, we created a new pipeline to generate large-scale synthetic ultrasound sequences with a wide range of cardiac deformations. Two types of synthetic data were thus generated, with and without reverberation artifacts.
- In contrast to [24], we showed that the PWC-Net architecture has the potential to produce relevant results on ultrasound images thanks to an adapted transfer learning procedure. This allows a better generalization of the network and a significant improvement of the results on clinical data.
- We further improve the performance of this network on ultrasound data by modifying its architecture to enhance its multi-scale analysis capability.
- We performed a thorough study of several temporal strategies that can be used to improve results during both the training and inference phases.
- We conducted the first study on the generalization of deep learning algorithms for motion estimation in echocardiography using a multi-center, multi-vendor and multi-disease dataset of real patients.

The interest of all contributions was carefully assessed both *in-silico* and *in-vivo* through standard geometric metrics and clinical indices.

II. Methods

A. Synthetic dataset for relevant transfer learning

Two recent studies have shown that supervised DL techniques can learn from synthetic ultrasound sequences to improve motion estimation on *in-vitro* [2] and *in-vivo* data [24]. In this context, the realism of synthetic image sequences is key for improving the performance of DL models. In both studies, a physical simulator was used to generate synthetic data and special care was taken to define a realistic medium from acoustic scatterers. Besides the realism of the ultrasound image, the motion must also be realistic. In [25], [26], the motion field was generated through a bio-mechanical personalized simulation. The personalization operation remains tedious, and currently limits the deployment of such scheme to small dataset (*i.e.* number of patients lower than 10 with the same kind of heart motion) with synthetic myocardial deformations that remain low as compared to reported normality ranges (*e.g.* simulated peak systolic longitudinal strain lower than 10% instead of 20% in real cases). In this paper, we proposed a dedicated simulation strategy to tackle this issue, and augment the database with diverse ranges on motions, cardiac geometries and image quality.

1) Overall strategy: Our overall strategy builds upon the same core concepts as in our previous papers [25], [26]. A schematic figure showing the workflow of the simulated pipeline is given in the supplementary materials. Clinical apical four-chamber B-mode recordings (called as template in the sequel) were used to simulate sequences with realistic tissue texture. For each frame of the template sequence, a scatter map was computed and fed to a physical simulator to produce the corresponding synthetic B-mode image. The scatterer maps were composed of two types of elements: the background and the myocardial scatterers. The full scatterers were distributed within the sector of the first frame according to a uniform random distribution. A density of 10 per square wavelength was chosen to ensure realistic speckle statistics. To avoid flickering effects, the background scatterers were kept

motionless. To mimic the local echogenicity of the recorded model, the local intensities I_m of the actual B-mode images were used to calculate the reflection coefficients RC_m of the scatterers, *i.e.* $RC_m = (I_m/255)^{(1/\gamma)} \cdot \mathcal{N}(0, 1)$, where $\mathcal{N}(\cdot)$ is the normal distribution, and γ is a constant for gamma compression. The myocardial scatterers were selected on the first simulated frame using manual annotations. The positions of these scatterers were then computed for each B-mode frame of the simulated sequence using the strategy described at the end of this section. The reflection coefficients of these scatterers were kept constant to maintain the speckle texture throughout the cardiac cycle. The final scatterers were obtained by combining the background and myocardial scatterers using the same scheme as in [25]. This strategy allows a smooth transition at the myocardial borders. Finally, a homemade open-source software called SIMUS from the MUST Matlab ultrasound toolbox[2] [27] was used to generate the synthetic ultrasound data. Each B-mode frame was generated by transmitting 128 focused beams, regardless of the acquired sector width (ranging from 60 to 90 degrees). In addition, the focal point was automatically chosen for each patient to be equal to half of the total acquired depth (ranging from 11 to 20 cm). The synthetic signals generated by SIMUS were demodulated to obtain IQ signals. The I/Q signals were beamformed using a delay-and-sum technique to obtain B-mode images [28].

2) Template image sequences: The template cine loops used in our simulation pipeline come from the CAMUS open access dataset which consists of exams from 500 patients acquired in clinical routine from the University Hospital of St-Etienne (France) and using a GE system [29]. This dataset was built without any specific image quality or patient selection criteria to match the heterogeneity of texture, shape and cardiac motions seen in clinical routine. We selected a subset of 100 apical four-chamber sequences, where the ultrasound machine settings were adjusted to scan the myocardium. The same probe settings used to acquire the CAMUS dataset were simulated: a 2.5 MHz 64-elements cardiac phased array.

3) Synthetic myocardial motion field: Endocardial and epicardial borders were delineated manually on the template sequences to obtain myocardial ROIs over the entire cardiac cycle. Time-varying surface meshes were generated for each of these ROIs following the resampling scheme given in Fig. 1. Specifically, the base of the left ventricle was defined by the segment linking the two extreme endocardial points. The apex was defined as the furthest point from the base in the epicardial contour. 36 points were then evenly distributed over the epicardial contour: 18 on the septum, and 18 on the lateral wall. Intramyocardial perpendicular segments were then drawn from these epicardial points to join the epicardial and endocardial contours. Each intramyocardial segment contained 5 evenly distributed points. This resampling scheme meshed the myocardium with 180 points (36 longitudinal $\times$ 5 radial) and 280 triangle cells. For each simulation, a set of points was randomly distributed over the myocardial mesh at end-diastole. Each of these points was then propagated over the full sequence by interpolating the displacements of the corresponding cell. This simple procedure allowed us to compute the temporal trajectory of any point belonging to the myocardium. The resulting synthetic myocardial motion field does not correspond to the actual motion field, which is not the purpose here. The interest of this procedure is to efficiently generate a wide variety of cardiac motions/deformations that are realistic enough to serve as a relevant data augmentation for DL methods.

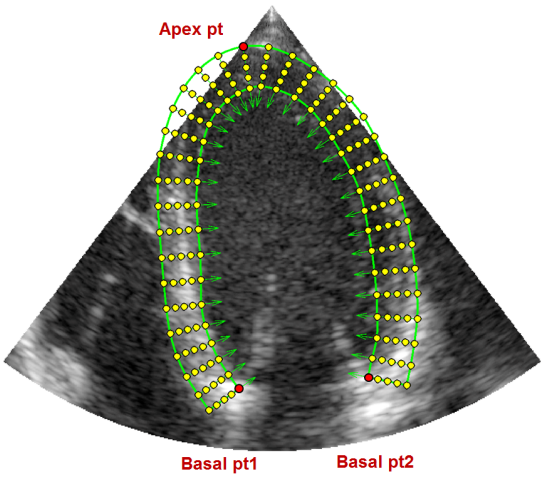


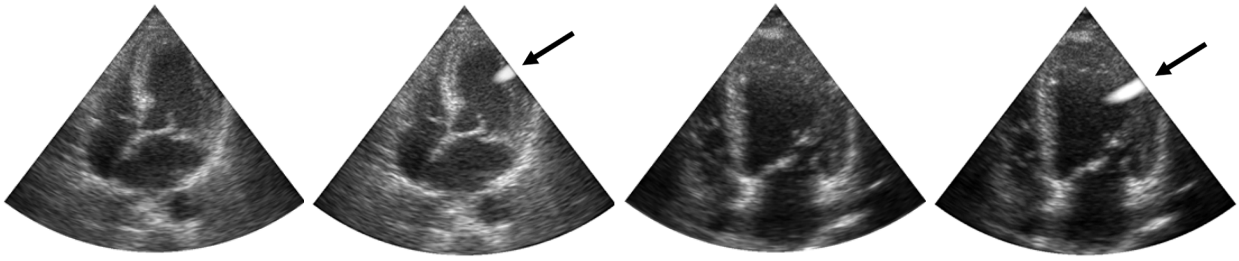
Fig. 1: Illustration of the resampling scheme used to generate a myocardial mesh (yellow nodes) from the corresponding segmentation mask (green lines).

4) Reverberation artifacts: We incorporated reverberation artifacts into our synthetic dataset to challenge the network during training. Specifically, we placed scatterers near the mid-anterolateral wall with high reflection coefficients relative to their neighbors. The position and amplitude of these scatterers remained constant throughout the cardiac cycle. This simple strategy leads to stationary saturated areas in the simulated B-mode images to emulate reverberation artifacts that may come from the ribs, as shown in Fig 2. Real reverberation artifacts may have other characteristics such as multiple reverberation structures and clutter noise, but these are not taken into account in this simulation.

B. Optimization of PWC-Net for echocardiography

1) Overall architecture: PWC-Net is one of the most efficient DL networks for dense motion estimation between two frames [12]. This network borrows the concept of a multi-resolution pyramidal structure to standard image tracking algorithms. Motion is estimated from the coarsest to the most detailed spatial resolution. A pyramid of seven levels with shared weights downsamples successively the features maps by half. Input images are processed separately. A normalized cross-correlation between the feature map of the first image and the second image warped by the previous estimated flow is then computed. This operation named *cost volume* performs patch comparisons between two feature maps for a range of displacements. The cost volume, the feature map from the first image, the upsampled estimated flow obtained at the

[2] www.biomecardio.com/MUST



(a) (b) (c) (d)

Fig. 2: Synthetic ultrasound images simulated from the proposed pipeline with (b, d) and without (a, c) reverberation artifacts for two different patients. The reverberation artifacts are identified by arrows.

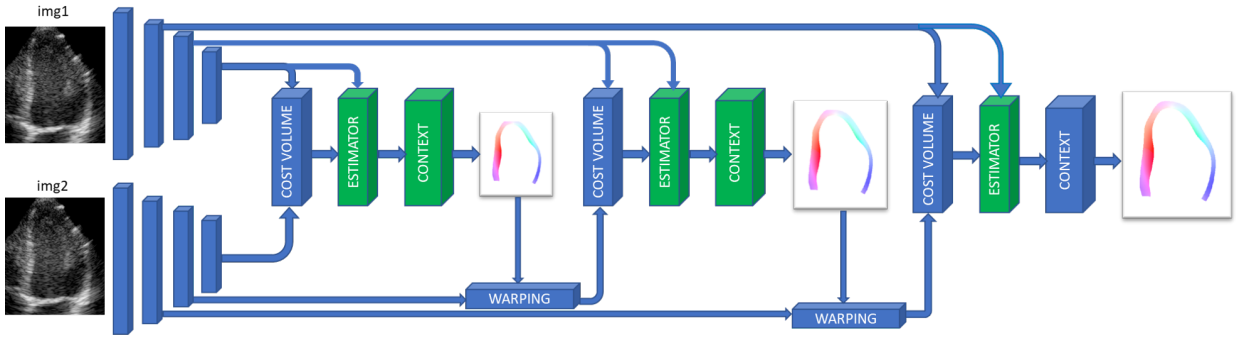


Fig. 3: Schematic view of our customized PWC-Net illustrated with a 4-level pyramid. The two input images are initially processed separately to extract the features, then the displacement fields are estimated in a coarse-to-fine manner (see Section II-B.1 for more details). The sub-networks modified as described in Sec. II-B.2 are displayed in green.

previous level and the upsampled feature map are used as input in a Convolutional Neural Sub-Network (CNSN), referred to as *estimator*. This CNSN is in charge of predicting a dense displacement map. The steps previously described are iterated until obtaining a displacement field with a quarter of the size of the initial input image. This information is then provided as input to another CNSN, referred to as *context*, to improve the accuracy of the estimated flow. This is done by adding the previously estimated flow with the output of a branch involving dilated convolutions to reinforce the receptive field. Finally, a bilinear interpolation upsamples the final flow to output a displacement map of the same size as the input image. The parameters of this network are optimized through a multi-scale loss function. This function computes the distance between the intermediate estimated flows and the corresponding scaled ground truths.

2) Proposed architecture: The overall architecture of our customized PWC-Net is given in Fig. 3 and 4. The modifications we made from the original architecture are all displayed in green. Based on the observation that multi-scale analysis has proven to be efficient for motion estimation in ultrasound [30], we first added a contextual sub-network at each resolution level of the network (context blocks in Fig. 3). In addition, the tracking of speckle patterns whose shapes can evolve between two consecutive frames make the motion estimation task particularly difficult in ultrasound. For this reason, we decided to reinforce the capacity of the network to extract relevant information by modifying each estimator sub-network

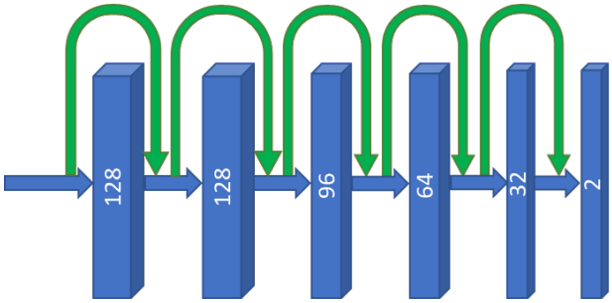


Fig. 4: Illustration of the estimator sub-network used in our customized PWC-Net with the added skip connections in green.

as illustrated in Fig. 4. These modifications correspond to skip connections concatenated to the output of each convolutional layer. The interest of these connections is twofold: *i)* since the PWC-Net architecture is deep, they limit the phenomenon of vanishing gradient; *ii)* the inputs of each convolutional layer are composed by the concatenation of the input and the outputs of the previous layer, leading to richer information sources. Similar to our intuition, Densenet connections were evaluated in [12], which improved the results by 5% but also increased the execution time up to 40%. Therefore, the authors leave the choice of using these connections according to the targeted objectives. The unlabeled blocks in Fig. 3 represent the pyramidal feature extractors described in Sec. II-B.1 and whose implementation details are described in Sec. III-C.1.

These feature extractors correspond to the ones proposed in the original PWC-Net implementation and do not involve any skip or residual connections.

3) Transfer learning strategy: In contrast to [24], we propose to keep the transfer learning strategy in order to strengthen the generalization capacity of the derived method, which should improve the results on clinical data. The specialization of the network from natural images to ultrasound was performed through different key steps. For adapting the proposed customized PWC-Net to gray level images, we first trained our network on a set of natural image pairs taken from the synthetic FlyingChairs2D and FlyingThings3D datasets [12]. Details on these datasets are given in Sec. III-A.1. Once the network has been learned on this first dataset, two different transfer learning procedures were performed on simulated ultrasound images with the same weight regularization as the initial training and without freezing any layer. The first transfer used an open access dataset [25] for the purpose of comparison with [24]. A second transfer was made using the same open access dataset extended with a new simulated ultrasound dataset based on the pipeline described in Sec. II-A. The properties of each synthetic dataset are provided in Sec. III-A.2. Both transfers used the same learning rate value ($\lambda = 1e^{-4}$) and the same progressive decay strategy as in the training on natural images, to ensure an efficient transfer to images of a different nature.

4) Temporal augmentation strategy: Different motion amplitudes between simulated and real data can worsen the performance of DL networks during inference. In addition, simulated data may be biased by certain types of motions and may not represent the variety of real movements, whether healthy or pathological. For addressing these issues, two temporal data augmentation strategies were combined during the training phase. First, to double the dataset with realistic movements of the ultrasound speckle, the pairs of forward frames with reference field ($t \rightarrow t+1$) were also presented to the network in the backward direction ($t+1 \rightarrow t$). In addition, rather than using only consecutive frames, we also provided image pairs separated by several frames to increase the amplitude of motion and the levels of speckle decorrelation seen by the network during training.

5) Composition inference strategy: The speckle motion pattern was assumed to be consistent for an image pair $I_1 I_2$ in the forward ($I_1 \rightarrow I_2$) and backward ($I_2 \rightarrow I_1$) directions. This forward-backward composition consistency was exploited during inference to still improve the performance of the network. In particular, each motion estimation between two consecutive B-mode frames was performed as follows. The forward motion between I_1 and I_2 (F_f) was first computed and used to propagate the myocardial points. The backward motion field (F_b) was then computed at these coordinates. In an ideal case, the composition of these two displacement fields should return the identity transformation. To respect that constraint, we averaged the forward F_f and backward $-F_b$ displacement fields to compute the final motion estimation.

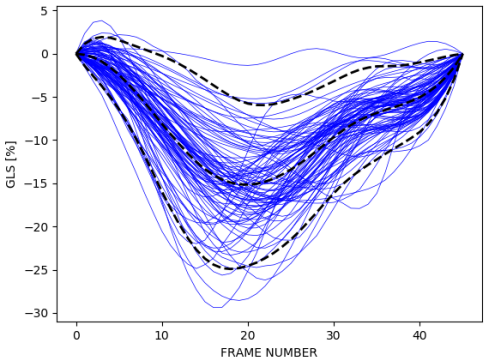


Fig. 5: Evolution of GLS as a function of time for all simulations. The mean and the limits of agreement are represented in black.

III. Experiments

A. Datasets

1) Synthetic natural datasets used for training: As described in Sec. II-B.3, our network was first trained to model motion estimation from natural synthetic images. To this aim, we used two public datasets consisting of image pairs with the corresponding dense displacement field on the entire image. The FlyingChairs dataset is composed of scenery images over which chairs with random orientations are over-imposed [8]. Random affine transformations were applied to the background and the chairs in the foreground. The FlyingThings3D dataset consists of images created from a mix of randomly textured 3D flying objects on a textured background [31]. The objects were randomly positioned in the image and then modified by geometrical transformations. Their motions correspond to randomized displacements along 3D linear trajectories. We removed image pairs with displacement amplitudes that were too large for what can be expected in echocardiography. Each image was expressed in grayscale format. This resulted to a synthetic dataset composed of 42,512 image pairs with the corresponding displacement fields.

2) Synthetic ultrasound datasets used for training and testing: We used an open access dataset of realistic 2D ultrasound sequences during the different transfer learning procedures [25]. This dataset is based on an electromechanical model of the heart which was combined with template cine loop recordings to simulate realistic ultrasound sequences. This approach relies on a personalization procedure which currently limits the heterogeneity of the dataset. This open access dataset is composed of 2D apical two-three-four chamber view sequences for seven vendors and five different motion patterns, including one healthy and four pathologies. This resulted in a dataset composed of 6,060 pairs of synthetic ultrasound images with the corresponding myocardial displacement fields.

To enrich this dataset, we developed a new simulation pipeline as described in Sec. II-A.1. Based on this new approach, we simulated 2D apical four chamber view sequences for 100 virtual patients from the CAMUS dataset. From Fig.

5, one can appreciate the rich variety and the realistic nature of the cardiac deformations present in the simulated dataset. The same template cine loops were used to simulate new sequences with reverberation artifacts. This increased the diversity of speckle patterns and can make the network less sensitive to this type of artifact. The resulting synthetic dataset is composed of 8,866 image pairs with the corresponding myocardial displacement fields. The full dataset is made available to the research community at `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`.

3) Clinical datasets used for testing: The open access CAMUS dataset described in Sec. II-A.2 was first used to create a 30-patient clinical dataset consisting of apical four chamber view sequences acquired from a GE system. These sequences were selected according to their quality (only good and medium image quality were included), also ensuring the whole left ventricular myocardium was included in the field of view. The original CAMUS dataset is provided with annotations only at end-diastole and end-systole. We therefore asked an expert cardiologist to extend these annotations to the entire cardiac cycle. This work was done through an in-house web annotation platform based on the Desk library[3] [32]. This resulted to clinical dataset composed of 1,443 image pairs with reference contours for both endocardial and epicardial borders.

In order to assess the generalization capacity of our approach, an auxiliary dataset composed of 2D apical four chamber view sequences from 30 new patients was collected at the University Hospital of Caen (France) within the regulation set by the local ethical committee. The images were acquired using Philips scanners. The same protocol as the one used for the CAMUS dataset was used to manually annotate the entire cardiac cycle. To ensure a representative range of left ventricle pathologies, five different groups equally distributed were selected, resulting in six patients from each group. The groups were defined by a diagnosis of aortic stenosis (AS), hypertrophic cardiomyopathy (HCM), ischemic heart failure (HF), non-ischemic HF and no disease. This resulted to an auxiliary clinical dataset composed of 1,536 image pairs with reference contours for both endocardial and epicardial borders.

B. Evaluated methods

As commercial applications like Tomtec and Echopac cannot be easily configured to modify the input contour to be tracked, we decided to assess the performance of the proposed network with the FFT-Xcorr block matching method [5], the Farnebäck optical flow method [6] and two DL methods, namely the PWC-Net [12] and the EchoPWC-Net [24]. Results from EchoPWC-Net and Farnebäck methods were taken directly from [24] due to difficulties in reproducing them.

1) FFT-Xcorr : Blockwise speckle tracking was implemented using standard FFT-based cross-correlations (FFT-Xcorr) on the B-mode images. The images were divided into subwindows with a 75% overlap. We detected small and large displacements with a multi-scale approach: displacement estimates were iteratively refined by decreasing the size of the subwindows (32×32, 16×16, and then 8×8). In contrast to [5], we computed the normalized cross-correlations from only two successive images (*i.e.*, no ensemble correlation). To obtain sub-pixel displacements, we used a parabolic fitting around the correlation peak. The estimated displacements were smoothed with an unsupervised robust spline smoother between two consecutive scales [33].

2) Farnebäck: This method is a traditional dense optical flow algorithm which is based on a pyramid of images at different resolution levels to track image points. A detailed description of the method and its parameters are given in [24].

3) EchoPWC-Net: This network is based on the PWC-Net architecture with several modifications to improve results on ultrasound data. The intuition behind these changes was to preserve information brought by speckle patterns by optimizing local variations. The authors also proposed to reinforce the resemblance of their solution with traditional speckle tracking approaches by integrating block matching aspects during the computation of the cost volume. The reader is referred to [24] for more details on this method.

C. Implementation details

Our experiments were realized using Keras 2.3.1 and Tensorflow 1.15 libraries. The modelling and experiments were conducted on a workstation with Ubuntu 20.10 operating system. The hardware consisted of an AMD Ryzen 9 3900X processor and an NVIDIA GeForce RTX 2080Ti GPU with 12 GB of memory. The runtime achieved in inference was 20 ms to estimate the myocardial motion between two consecutive frames.

1) Architecture parameters: For the feature extractor of each level of the pyramid, we used three convolutional layers, including a convolution with a stride of two to downsample the final feature maps. The number of filters per level is constant and was set to 16, 32, 64, 96, 128 and 196 from top to bottom. The search range of the cost volume was fixed to 4 pixels. The same context network as the one proposed in [12] was used at each resolution level, composed by a succession of dilated convolutions with factors of 1, 2, 4, 8, 16, 1 and 1 respectively. The total number of parameters of our network was 14M.

2) Training procedures: Our network was trained with the Adam optimizer and a batch size of 4 for all the experiments. The initial learning rate was set to $1e^{-4}$, with a halving schedule every 20% of the training after 40% of the total number of iterations. 200 epochs (150 hours) and 15 epochs (15 hours) were used during the training on the natural and ultrasound datasets, respectively. The weights were initialized following the He initialization scheme [34].

3) Data augmentation: Different data augmentation strategies were applied depending on the learning phase. During training on the synthetic dataset with natural images, the same data augmentation scheme as the one performed in [8] was carried out. A random crop of size 448×384 was used to

[3] `www.creatis.insa-lyon.fr/~valette/public/project/desk/`

select the images given as input to the network. In addition, geometric transformations (translations, scaling, flipping) and image alterations (brightness, Gaussian noise, contrast) were used. During the transfer learning phase, only image alterations (brightness and contrast) were used to respect the shapes and geometrical properties of the ultrasound images.

4) Loss: We used a multi-scale loss function defined as

$$L(\Theta) = \sum_{l=l_0}^{L} \alpha_l \sum_x |w_\theta^l(x) - w_{GT}^l(x)|_2 + \gamma|\Theta|_2, \quad (1)$$

with Θ being the parameters, x the inputs, $|.|_2$ the $L2$ norm and w^l the flow field at the lth pyramid level. The values of the α_l parameters correspond to the one proposed in [8]. The regularization factor γ was set to $4e-4$. During transfer learning, the optimization of this loss was restricted to the region where the reference dense displacement field was known, *i.e.* in the myocardial region.

D. Evaluation metrics

1) Geometrical metrics: To measure the accuracy of the estimated motion and the tracked contours of a given method, the endpoint error (EPE), the mean absolute distance (d_m) and the Hausdorff distance (d_H) were used. The endpoint error is defined as $EPE = \sqrt{(u-u_{gt})^2 + (v-v_{gt})^2}$ and corresponds to a distance measure between the estimated flow (u, v) and the ground truth (u_{gt}, v_{gt}) [35]. This metric was computed only in the myocardial region. d_m corresponds to the average distance between two contours while d_H measures the local maximum distance between the two contours. In our experiments, d_m and d_H were used to assess tracking quality by comparing the reference endocardial contours with the contours obtained by propagating the reference contour at the first frame with the successive motion estimations obtained with the different techniques. The reported d_m and d_H scores correspond to the average values computed from all the contours obtained on the same sequence except the one at the first frame.

2) Clinical metrics: We gauged the methods' performance with the Global Longitudinal Strain (GLS). There is currently no consensus on the way to compute it. Strain estimation from the epicardial border is generally avoided because of the proximity of the pericardium, which makes several areas difficult to segment. Because of a varying quality of the speckle pattern along the heart muscle, strain estimation from the myocardium is usually associated with a regularization term that affects the quality of the measurements. Finally, the strain estimation from the endocardium can be seen as the more robust to achieve thanks to a sufficient contrast between the blood pool and the myocardium. However, it is known that the computation of the GLS from the endocardium takes into account a part of the radial deformation [36]. In this study, we decided to use the endocardial contour to compute the longitudinal ventricular length L for each timestep t to estimate the Lagrangian strain, defined as:

$$S_L = \frac{L(t) - L(t_0)}{L(t_0)}, \quad (2)$$

where $L(t)$ stands for the longitudinal length at a given time t with t_0 corresponding to end-diastole. Moreover, to address the weakness of using the endocardial contour to calculate GLS, we use a complementary clinical index that provides information exclusively on longitudinal deformation, named Mitral Annular Plane Systolic Excursion (MAPSE) [37]. This index is computed in the same way as the GLS, with the difference that L denotes the distance between the apex and the mid-basal point of the endocardial border. Both GLS and MAPSE were computed for each time of the sequence. The peak-systolic strain, defined as the minimum between end-diastole and end-systole strain values, was finally calculated for both indices, for which we computed three metrics: the bias, the standard deviation σ and the mean absolute error (mae). GLS and MAPSE ground truths were derived from the expert manual annotation of the endocardial contour. For fair comparisons, the same endocardial reference contours at end-diastole were used as the initial contours from which the different tracking methods were applied. All reference contours were manually delineated prior to the application of any method, thus avoiding any bias issues. It is important to note that the quality of the results of our tracking depends on the initial contour that is provided as input by an operator. Particular attention was therefore paid for accurately segmenting the myocardium.

IV. Results

A. Simulation

1) Ablation study: We first conducted an ablation study to evaluate the influence of different design choices during the training and inference phases. We used for this purpose the open-access dataset of synthetic ultrasound sequences described in Sec. III-A.2. This dataset was divided into folds by vendor, of which five were used for training, one for validation and one for testing. During this experience, we trained our customized version of PWC-Net, referred as c-PWC-Net in the sequel, using a composition of the following temporal data augmentation strategies: *i)* using data in the forward direction only; *ii)* using data in the forward/backward directions and *iii)* using image pairs separated by one, two or three frames. During inference, the composition consistency described in Sec. II-B.5 was also investigated. The obtained results are reported in Table I. From these results, we can first observe the slight improvement (0.01 mm reduction of the EPE) brought by the composition consistency procedure. Regarding the training phase, the forward/backward and the spaced image pairs strategies further improved the EPE. The best configuration was obtained when combining the composition consistency in inference with the forward/backward data augmentation strategy during training using image pairs separated by one and two frames, the separation with more images leading to an unchanged average EPE of 0.04 ± 0.03 mm. This configuration was therefore employed for all the following experiments.

The importance of the modifications we brought to the original PWC-Net architecture was investigated through two complementary ablation studies. The first experiment was

TABLE I: Ablation study on the open access dataset [25] for the temporal data augmentation and inference strategies proposed in Sec. II-B. The proposed customized PWC-Net (c-PWC-Net) was trained using forward (+), forward-backward (+/-) and image pairs separated by n frames (0 meaning two consecutive images). In inference, results were computed with and without the composition consistency procedure given in Sec. II-B.5.

c-PWC-Net		**EPE**$\pm\sigma$
Training	Inference	mm.
+, 0	No composition	0.09 $\pm$ 0.07
+, 0	Composition	0.08 $\pm$ 0.07
+/-, 0	Composition	0.05 $\pm$ 0.04
+, 0, 1	Composition	0.05 $\pm$ 0.05
+/-, 0, 1	Composition	0.04 $\pm$ 0.04
+, 0, 1, 2	Composition	0.05 $\pm$ 0.03
+/-, 0, 1, 2	Composition	**0.04** $\pm$ **0.03**
+, 0, 1, 2, 3	Composition	**0.04** $\pm$ **0.03**
+/-, 0, 1, 2, 3	Composition	**0.04** $\pm$ **0.03**

realized on the same dataset as for the ablation study given in Table I. During this experiment, we trained c-PWC-Net using the same transfer learning strategy but for different versions of its architecture: *i)* without skip connection and contextual sub-networks; *ii)* with skip connection only; *iii)* with contextual sub-networks only and *iv)* with both of them (which corresponds to the proposed full architecture). From the results given in Table II, one can clearly see the importance of the skip connections and the contextual sub-networks, the use of the contextual sub-networks resulting in an improvement of 0.2 mm while its combination with skip connections allowed to divide by 2 the average EPE compared to the original architecture of PWC-Net (*i.e.* c-PWC-Net without skip connection and contextual sub-network). Contrary to the two previous ablation studies, we used the proposed synthetic dataset to study the influence of the contextual sub-networks. Indeed, this dataset presents more variety in terms of strain deformations and therefore seems to be better suited. During this experiment, we trained c-PWC-Net using the same transfer learning strategy and with the following conditions: *i)* with a contextual sub-network for each resolution, except the two lowest ones; *ii)* with a contextual sub-network for each resolution, except the lowest one; *iii)* with a contextual sub-network for all resolutions (*i.e.* 6 sub-networks in total). Results given in Table III clearly show the interest of adding a contextual sub-network at each resolution level, even for the lowest ones. The best network architecture, subsequently adopted, was therefore the combination of skip connections with contextual sub-networks for all resolutions.

2) Open access synthetic US dataset: Table IV shows the motion estimation performance of c-PWC-Net from the open access synthetic US dataset [25]. We used the same nomenclature as the one introduced in [24], by adding "-gray" in case of training performed on grayscale natural images, "-us" in case of training on the open access dataset and "-ft" in case of fine-tuning. For instance, PWC-Net-gray-usft refers to PWC-Net first trained on natural grayscale images, then on the synthetic

TABLE II: Ablation study performed on the open access dataset [25] for the architectural modifications given in Sec. II-B.2. The different networks were trained in the same conditions using the forward-backward strategy with image pairs separated by 0, 1, and 2 frames.

c-PWC-Net architecture		**EPE**$\pm\sigma$
Skip connections	Contextual sub-net.	mm.
✘	✘	0.08 $\pm$ 0.06
✔	✘	0.07 $\pm$ 0.06
✘	✔	0.06 $\pm$ 0.06
✔	✔	**0.04** $\pm$ **0.03**

TABLE III: Ablation study performed on the proposed synthetic dataset for the influence of the contextual sub-networks presented in Sec. II-B.2. The different networks were trained under the same conditions as for the other ablation studies. The column labeled Positions provides information about the presence of a contextual sub-network relative to the pyramid level (1 stands for the highest resolution, 6 to the lowest).

c-PWC-Net architecture		**EPE**$\pm\sigma$
Number of sub-net.	Positions	mm.
4	1, 2, 3, 4	0.10 $\pm$ 0.08
5	1, 2, 3, 4, 5	0.09 $\pm$ 0.07
6	1, 2, 3, 4, 5, 6	**0.07** $\pm$ **0.06**

US dataset of [25] through transfer learning. The Farnebäck, PWC-Net-gray-usft and EchoPWC-Net-us results were taken from [24]. Concerning c-PWC-Net-gray-usft, the first training performed on the gray-scaled FlyingChairs and FlyingThings datasets was performed by splitting the full set of images into training and validation sets. On the validation set, the network achieved an average EPE of 1.53 ± 5.28 pixels. The comparison of the results obtained by the PWC-Net-gray-usft method using either the transfer learning strategy given in [24] or the one we proposed in Sec. II-B.3 clearly illustrates the relevance of the choices we made. Indeed, using the same original PWC-Net architecture, our transfer learning strategy yields an overall improvement of 42% in the average EPE on the full dataset, from 0.14 mm to 0.08 mm. In addition, it appears that the architectural modifications we proposed further improve the results obtained by c-PWC-Net-gray-usft by reducing the average EPE by 0.02 mm. It is also worth noting that the two non-DL methods outperformed c-PWC-Net-gray, which gave notably bad results on the GE fold. This can be explained by the fact that c-PWC-Net-gray is a method trained only on simulated natural images. Its performance is naturally reduced on echocardiographic images. In the open access dataset of Alessandrini *et al.*, the simulated sequences from the GE vendor are found to be the most challenging in terms of image quality with the least defined speckle pattern, making the tracking task more difficult. This observation is confirmed by the fact that all evaluated methods score their worst on this fold. Finally, the best performing methods in this experiment were obtained by the two DL techniques EchoPWC-Net-us and c-PWC-Net-gray-usft which reached the same average EPE of 0.06 ± 0.05 mm.

TABLE IV: Results on the open access synthetic dataset [25]. The methods are compared on seven vendors in apical four chamber view. The metric used is the average endpoint error expressed in mm. The application of the Wilcoxon signed-rank test shows the statistical difference ($p < 0.0001$) of c-PWC-Net-gray-usft with the methods for which we have the results for all the patients (referred by *).

Methods	*ESAOTE*	*GE*	*HITACHI*	*PHILIPS*	*SIEMENS*	*TOSHIBA*	*SAMSUNG*
	mm.	mm.	mm.	mm.	mm.	mm.	mm.
Farnebäck [38]	0.08 (0.06)	0.09 (0.07)	**0.06 (0.04)**	0.08 (0.06)	0.06 (0.05)	0.07 (0.05)	0.07 (0.05)
FFT-Xcorr [5]*	0.10 (0.08)	0.14 (0.11)	0.11 (0.08)	0.09 (0.07)	0.09 (0.08)	0.09 (0.07)	0.09 (0.07)
PWC-Net-gray-usft [24]	0.14 (0.10)	0.17 (0.12)	0.13 (0.09)	0.14 (0.10)	0.14 (0.10)	0.14 (0.11)	0.13 (0.09)
EchoPWC-Net-us [24]	**0.07 (0.06)**	**0.07 (0.06)**	**0.06 (0.04)**	**0.06 (0.05)**	0.06 (0.05)	**0.06 (0.04)**	**0.05 (0.04)**
PWC-Net-gray-usft (ours)*	0.08 (0.07)	0.10 (0.07)	0.07 (0.04)	0.09 (0.06)	0.08 (0.06)	0.07 (0.05)	0.08 (0.06)
c-PWC-Net-gray*	0.15 (0.13)	0.34 (0.72)	0.11 (0.08)	0.12 (0.10)	0.09 (0.08)	0.12 (0.10)	0.13 (0.08)
c-PWC-Net-gray-usft	**0.07 (0.06)**	0.08 (0.07)	**0.06 (0.04)**	0.07 (0.06)	**0.04 (0.03)**	**0.06 (0.04)**	0.07 (0.05)

3) Proposed synthetic US dataset: We conducted experiments to assess the added value of the proposed synthetic ultrasound dataset described in Sec. II-A. The corresponding results are given in Tables V and VI. In these experiments, the training dataset was composed of the open access dataset proposed in [25] augmented with data from 60 virtual patients, while the validation and testing datasets were composed of the data from the 10 and 30 remaining virtual patients. Patients were randomly selected to create the different datasets. Several trainings were performed by varying the number of added virtual patients (from 20 to 60) and by including or not the data with reverberation artifacts for the same patients. In these tables, the different experiments are named according to the number of added patients, with an A appended if the simulated patients were included with and without artefacts. For instance, c-PWC-Net-20A refers to c-PWC-Net trained using the open access dataset augmented with 20 virtual patients with and without reverberation artifacts. We compared the performance of the c-PWC-Net with the FFT-Xcorr block-matching approach. In these experiments, c-PWC-Net outperformed FFT-Xcorr for most of the metrics and training configurations.

In terms of geometric scores, it is worth noting the continuous improvement in overall scores with the increasing amount of synthetic data both in terms of mean and standard deviation, from 1.53 ± 1.12 mm to 0.73 ± 0.49 mm for the d_m metric and from 0.14 ± 0.11 mm to 0.07 ± 0.06 mm for the EPE computed on the testing dataset without artifact. The same trend can be observed with the computation of the cumulative EPE between ED and ES, *i.e.* an improvement of c-PWC-Net from 2.66 ± 1.59 mm to 1.20 ± 0.67 mm when adding the 60 virtual patients. The same conclusions can be drawn on the data with artifacts, with an improvement from 1.64 ± 1.16 mm to 0.79 ± 0.53 mm for the d_m metric and from 0.16 ± 0.13 mm to 0.08 ± 0.06 mm for the EPE. The positive impact of including synthetic data with and without reverberation artifacts appeared systematically and for all metrics in Table V.

For the clinical scores given in Table VI, the same trends as for the geometric metrics can be drawn, with an overall improvement of the different scores with the inclusion of more synthetic data. Concerning the bias and standard deviations for the two clinical indices, the main improvement occurred after adding the first 20 patients, from 2.95 ± 3.00 % to 0.57 ± 1.73 % for the GLS and from 4.72 ± 3.54 % to 1.45 ± 2.47 % for the MAPSE on the testing dataset without artifact. The addition of more data resulted in a stagnation of the bias, but a decrease in the standard deviation, leading to an overall improvement of both indices. The same happened for the testing dataset with reverberation artifacts. For both clinical indices, the addition of the 60 simulated patients (with or without artifacts) allowed a final improvement around 80% for the bias and 65% for the mae.

B. Clinical data

1) Real patients from the CAMUS dataset: The performance of c-PWC-Net was assessed on the clinical data described in Sec. III-A.3. FFT-Xcorr, c-PWC-Net and c-PWC-Net-60A, which performed best on the synthetic ultrasound dataset, were compared in this experiment. The comparison of our method with EchoPWC-Net would have been of interest. Unfortunately this is impossible since neither the trained model nor the test dataset were made publicly available by the authors. The geometric scores are given in Table VII. Since real motions are not known, only d_m and d_H metrics were computed from the tracked contours. c-PWC-Net outperformed the FFT-Xcorr block-matching method, with an improvement of 0.72 mm for d_H metric. Another interesting point is the improvement of our model brought by the proposed synthetic dataset, resulting in an overall performance of 1.86 ± 1.05 mm for d_m and 3.81 ± 1.18 mm for d_H. Table VIII lists the scores obtained for the GLS and MAPSE clinical indices. The same trends as observed for the geometric metrics can also be drawn. Indeed, our DL solution outperformed FFT-Xcorr while the use of the synthetic dataset significantly improved the different clinical scores, with a mae from 4.29 ± 2.84 % to 2.55 ± 2.08 % for the GLS and from 4.19 ± 2.78 % to 2.62 ± 2.09 % for the MAPSE. As a complement, a correlation plot between the estimated GLS and the ones from the ground truth is given in the supplementary materials. A correlation coefficient of 0.77 was achieved, demonstrating the capacity of our method to reproduce manual annotations with good fidelity. It is also important to note that our correlation score was not as good as

TABLE V: Geometric results on the open access synthetic dataset [25] complemented with the proposed simulated database. The p-value computed on the EPE with the Mann-Whitney U rank test between FFT-Xcorr and c-PWC-Net-60A was equal to $6e^{-6}$, proving the statistical difference between these two methods.

Methods*	***Simulations***			***Artifacts***		
	EPE$\pm\sigma$	$d_m \pm \sigma$	$d_H \pm \sigma$	EPE$\pm\sigma$	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.	mm.	mm.	mm.	mm.
FFT-Xcorr	0.26 ± 0.18	1.63 ± 0.97	4.64 ± 1.88	0.27 ± 0.19	1.86 ± 1.17	5.02 ± 2.02
c-PWC-Net-0	0.14 ± 0.11	1.53 ± 1.12	3.76 ± 1.36	0.16 ± 0.13	1.64 ± 1.16	4.07 ± 1.51
c-PWC-Net-20	0.09 ± 0.07	0.87 ± 0.59	2.37 ± 0.68	0.10 ± 0.09	1.06 ± 0.79	2.68 ± 0.84
c-PWC-Net-20A	0.08 ± 0.06	0.84 ± 0.57	2.32 ± 0.66	0.09 ± 0.07	0.90 ± 0.61	2.30 ± 0.64
c-PWC-Net-40	0.08 ± 0.06	0.79 ± 0.53	2.12 ± 0.61	0.09 ± 0.08	0.96 ± 0.70	2.36 ± 0.75
c-PWC-Net-40A	0.08 ± 0.06	0.78 ± 0.52	2.05 ± 0.56	**0.08 ± 0.06**	0.84 ± 0.57	2.13 ± 0.60
c-PWC-Net-60	0.08 ± 0.06	**0.73 ± 0.49**	**1.98 ± 0.54**	0.09 ± 0.07	0.92 ± 0.70	2.34 ± 0.74
c-PWC-Net-60A	**0.07 ± 0.06**	**0.73 ± 0.49**	2.03 ± 0.56	**0.08 ± 0.06**	**0.79 ± 0.53**	**2.08 ± 0.60**

TABLE VI: Clinical metrics on the open access synthetic dataset [25] complemented with the proposed simulated database.

Methods	***Simulations***				***Artifacts***			
	GLS		*MAPSE*		*GLS*		*MAPSE*	
	bias$\pm\sigma$	mae$\pm\sigma$	bias$\pm\sigma$	mae$\pm\sigma$	bias$\pm\sigma$	mae$\pm\sigma$	bias$\pm\sigma$	mae$\pm\sigma$
	%.	%.	%	%	%.	%.	%	%
FFT-Xcorr	4.89 ± 1.69	4.89 ± 1.69	2.77 ± 1.25	2.77 ± 1.25	6.01 ± 2.12	6.01 ± 2.12	3.20 ± 1.24	3.20 ± 1.24
c-PWC-Net-0	2.95 ± 3.00	3.40 ± 2.46	3.26 ± 2.13	3.30 ± 2.06	4.72 ± 3.54	4.94 ± 3.21	4.34 ± 2.45	4.39 ± 2.35
c-PWC-Net-20	0.57 ± 1.73	1.56 ± 0.90	0.56 ± 1.36	1.17 ± 0.88	1.45 ± 2.47	2.43 ± 1.47	0.98 ± 1.45	1.48 ± 0.91
c-PWC-Net-20A	0.48 ± 1.72	1.54 ± 0.88	0.43 ± 1.24	1.03 ± 0.79	0.70 ± 1.86	1.70 ± 0.99	0.54 ± 1.20	1.09 ± 0.72
c-PWC-Net-40	0.68 ± 1.47	1.35 ± 0.87	0.69 ± 1.13	1.08 ± 0.74	1.57 ± 2.11	2.29 ± 1.24	1.10 ± 1.33	1.47 ± 0.88
c-PWC-Net-40A	0.83 ± 1.30	1.26 ± 0.87	0.60 ± 1.00	0.93 ± 0.70	0.97 ± 1.60	1.57 ± 1.00	0.64 ± 1.03	1.02 ± 0.64
c-PWC-Net-60	**0.66 ± 1.31**	**1.20 ± 0.82**	0.60 ± 0.89	0.89 ± 0.60	1.50 ± 1.89	2.08 ± 1.20	1.00 ± 1.06	1.25 ± 0.73
c-PWC-Net-60A	0.59 ± 1.44	1.27 ± 0.87	**0.55 ± 0.90**	**0.85 ± 0.62**	**0.71 ± 1.55**	**1.42 ± 0.92**	**0.63 ± 0.92**	**0.95 ± 0.58**

TABLE VII: Geometric results obtained on a subset of the CAMUS dataset composed of 30 real patients acquired with a GE system.

Methods	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.
FFT-Xcorr	2.27 ± 1.30	5.36 ± 2.07
c-PWC-Net	2.22 ± 1.34	4.64 ± 1.62
c-PWC-Net-60A	**1.86 ± 1.05**	**3.81 ± 1.18**

the one obtained in the recent study of [24]. This can be explained by the fact that the authors compare their method with the results obtained by commercial methods usually based on conventional speckle tracking algorithms. Using manual expert annotations as references can be more difficult because they are not based solely on image information and incorporate a variety of complex information that may be subjective. Finally, we observed that c-PWC-Net-60A has stable performances even for images of lower quality (a dedicated table is provided in the supplementary materials). This can be explained by the fact that the training dataset included a large range of image quality, and is encouraging as for the generalization capability of our algorithm.

2) Real patients from the auxiliary dataset: The generalization capacity of c-PWC-Net-60A was assessed using the auxiliary dataset described in Sec. III-A.3. This dataset contains echocardiographic sequences acquired exclusively from another hospital with a system from a different vendor than the one used to create the synthetic data. Moreover, this dataset was not used to generate new synthetic cases, so our algorithm never integrated this new type of data during its learning phase. From Table IX, it can first be observed that the geometric scores remain unchanged, with a mean value of 1.81 ± 1.11 mm for d_m and 3.45 ± 1.11 mm for d_H. It is also interesting to see that the quality of the tracking is homogeneous with respect to the type of pathology, with a variability of 0.58 mm for d_m and 0.88 for d_H.

Table X shows that the same trends are true for the clinical scores. Indeed, c-PWC-Net-60A obtained very similar results on the Philips dataset compared to the GE one, with a mae of 2.89 ± 2.08 % and 2.86 ± 1.88 % for the GLS and MAPSE, respectively. These results are also consistent between the different pathological groups. A correlation plot between the estimated GLS and the reference ones is also given in the supplementary materials. Interestingly, a correlation coefficient of 0.93 was obtained, which can be explained by an overall better image quality in the Philips dataset.

During the manual contouring of the testing datasets, the points used to define the reference contours were selected independently from one frame to another by the expert cardiologist. We therefore have no reference point tracked over the cardiac cycle, which prevents us from computing the reference regional strain. Nevertheless, we conducted an additional experience to assess the spatial distribution of the distance errors between the reference and the estimated contours. In

TABLE VIII: Clinical results obtained on a subset of the CAMUS dataset composed of 30 real patients acquired with a GE system.

Methods	*GLS*		*MAPSE*	
	bias$\pm\sigma$	**mae**$\pm\sigma$	**bias**$\pm\sigma$	**mae**$\pm\sigma$
	%.	%.	%.	%.
FFT-Xcorr	7.35 $\pm$ 3.42	7.35 $\pm$ 3.42	5.66 $\pm$ 3.43	5.66 $\pm$ 3.43
c-PWC-Net	3.96 $\pm$ 3.30	4.29 $\pm$ 2.84	3.90 $\pm$ 3.19	4.19 $\pm$ 2.78
c-PWC-Net-60A	**1.85** $\pm$ **2.73**	**2.55** $\pm$ **2.08**	**1.83** $\pm$ **2.83**	**2.62** $\pm$ **2.09**

TABLE IX: Geometric results obtained with the c-PWC-Net-60A method on an auxiliary dataset composed of 30 real patients acquired with a Philips system.

Philips dataset	$d_m \pm \sigma$	$d_H \pm \sigma$
	mm.	mm.
Full dataset (#30)	1.81 $\pm$ 1.11	3.45 $\pm$ 1.11
Aortic Stenosis (#6)	1.72 $\pm$ 1.11	3.24 $\pm$ 1.02
Hypertrophic Cardiomyopathy (#6)	2.15 $\pm$ 1.26	3.91 $\pm$ 1.36
Ischemic (#6)	1.67 $\pm$ 1.08	3.38 $\pm$ 1.09
Non Ischemic (#6)	1.57 $\pm$ 0.95	3.03 $\pm$ 0.96
Normal (#6)	1.93 $\pm$ 1.14	3.69 $\pm$ 1.14

particular, the orientation of the long axis was first extracted from the first frame of the processed sequence. For each contour of the sequence, apical points were defined as the intersection between the given contour and the line passing by the mid-basal point with the same orientation as the one computed on the first frame. A normalized parametrization for each contour was then computed, where 0 corresponds to the basal anterolateral point, 0.5 corresponds to the apex point and 1 corresponds to the basal inferoseptum point. Mean absolute distance between the reference and the estimated contours were finally computed along the parametric axis. Fig. 6 displays the results obtained from the auxiliary dataset. Each curve corresponds to one of the 30 evaluated patients. On this figure, we can see that the errors are relatively homogeneous on each side of the myocardium, with slightly higher average values for the lateral side and at the apex (*i.e.* a mean error of 1.4 mm on the septal side and 1.9 mm on the lateral side).

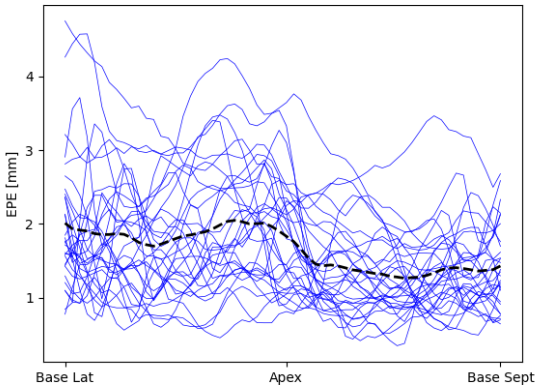


Fig. 6: Evolution of the distance error along the endocardium between the reference contours and the estimated ones using c-PWC-Net-60A. Each blue contour corresponds to the mean error computed over the cardiac cycle for one patient. The mean curve is represented in black.

V. Discussion

A. A new open access simulated ultrasound dataset

The open access dataset of [25] was generated with synthetic deformation in lower ranges than normal strain values. Moreover, the complex personalization procedure in [25] limited the variability of geometries and motion types that can be simulated. This limits the relevance of using this dataset alone for DL training. Therefore, we designed a solely image-based pipeline, bypassing the need for an electromechanical model. This allowed to simulate many cases from B-mode template cine loops on which myocardial contours were manually annotated to generate synthetic motion fields. As illustrated in Fig. 5, global deformation ranges with our simulation method match those of real sequences, although at a finer local scale disparities can occur. It is thus important to exploit this synthetic dataset with care. Indeed, although the corresponding global clinical indices seem relevant, the way in which the baseline myocardial motion was generated does not allow our dataset to be used to evaluate motion estimation algorithms. The accuracy of the myocardial motion pattern is necessarily limited and the proposed simulation pipeline should be viewed as a synthetic ultrasound sequence generator for data augmentation purposes only. With this in mind, we used this simulation pipeline to augment the existing open-access dataset of [25] and produced a more diverse synthetic dataset. Tables V and VI showed the interest of our simulation strategy, with an improvement of the geometric and clinical scores when increasing the number of virtual patients in the training dataset. Specifically, the geometrical errors were mostly reduced by a factor of two thanks to the addition of the 60 simulated patients. Our pipeline can also simulate physical artifacts to improve the robustness of DL to these application-specific sources of noise. We focused in this study on reverberation artifacts. As shown in Table VI, incorporating these artifacts in the training dataset significantly improved both the GLS and MAPSE scores, if the same type of artifact is present in the testing dataset. This validates the relevance of generating synthetic images with artifacts as a data augmentation procedure. The full set of the simulated

TABLE X: Clinical results obtained with the c-PWC-Net-60A method on an auxiliary dataset composed of 30 real patients acquired with a Philips system.

Philips dataset	***GLS***		***MAPSE***	
	bias$\pm\sigma$	**mae**$\pm\sigma$	**bias**$\pm\sigma$	**mae**$\pm\sigma$
	%.	%.	%.	%.
Full dataset (#30)	2.85 $\pm$ 2.14	2.89 $\pm$ 2.08	2.74 $\pm$ 2.05	2.86 $\pm$ 1.88
Aortic Stenosis (#6)	2.85 $\pm$ 2.14	2.85 $\pm$ 2.14	2.22 $\pm$ 1.91	2.46 $\pm$ 1.52
Hypertrophic Cardiomyopathy (#6)	3.33 $\pm$ 2.26	3.33 $\pm$ 2.26	3.51 $\pm$ 2.27	3.51 $\pm$ 2.27
Ischemic (#6)	2.48 $\pm$ 1.59	2.50 $\pm$ 1.56	2.98 $\pm$ 1.29	2.98 $\pm$ 1.29
Non Ischemic (#6)	1.82 $\pm$ 1.92	2.01 $\pm$ 1.67	2.51 $\pm$ 1.77	2.51 $\pm$ 1.77
Normal (#6)	3.75 $\pm$ 2.84	3.75 $\pm$ 2.84	2.51 $\pm$ 3.08	2.85 $\pm$ 2.69

data is made publicly available. We claim that the access to this synthetic dataset in addition to the one proposed in [25] will provide valuable and complementary tools for the research community. To better evaluate the quality of the simulated dataset, several videos of synthetic sequences are available at: `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`. Finally, it is worth noting that the manual annotation step involved in the current version of the simulation pipeline can be time consuming and is the main bottleneck to automatically deploy our solution on more than 100 patients regardless of the type of view during acquisition. This is the reason why we limited our study to 100 patients acquired on the apical 4-chamber view. In the near future, we plan to work on a fully automated and controlled deployment of our simulation pipeline to be able to generate larger and richer synthetic databases in terms of diversity of cases, pathologies and acquisition view.

B. Interest of the training/inference strategies

The use of data augmentation is key for DL-based methods. It thus appeared appealing to investigate dedicated strategies for tracking in ultrasound. Table I shows a reduction of the mean and standard deviation of the EPE by a factor of 2 when using forward/backward and spaced image pairs. This can be explained by the increase in motion diversity of the dataset while preserving the decorrelation of the associated speckle pattern. In parallel, we proposed a strategy for inference that incorporates temporal consistency. Although it provided a 11% improvement in the mean EPE, this procedure doubles the inference time, limiting its application to scenarios for which computation time is not a strong constraint.

C. Efficiency of the proposed transfer learning solution

We evaluated the relevance of our transfer learning solution on both simulated and clinical data. We first benchmarked our network against EchoPWC-Net [24] on the dataset of [25]. As illustrated in Table IV, the results obtained by the two networks are similar, despite different choices to extend the original PWC-Net architecture for ultrasound image processing. The authors of EchoPWC-Net obtained their best results by learning directly from simulated echocardiographic images, without any transfer learning. In contrast, we opted for a transfer learning approach starting first on natural images, before transferring to ultrasound data. Contrary to [24], we assumed that first confronting the network with a wide variety of images and motion types would enable a better generalization ability and avoid overfitting. In [24], transfer learning was evaluated but performed poorly. There may be several reasons for this, including the limited variability in geometry and motion types of the synthetic dataset used for training and the lack of learning on the ultrasound synthetic data due to a low learning rate. In our case, the learning rate was left unchanged between initial and transfer learning phases. The interest of our transfer learning approach was further validated by the results obtained on clinical data. Indeed, Tables VII and VIII showed that our method still significantly outperformed the state-of-the-art FFT-Xcorr block-matching method, with 18% and 29% improvements in the d_m and d_H scores, respectively. This improvement was even bigger for the clinical indices, with a reduction of the mae from 7.3% to 2.5% for the GLS and from from 5.7% to 2.6% for the MAPSE. It would have been interesting to compare our method with EchoPWC-Net on clinical data. Unfortunately, this was impossible as neither the testing dataset nor the commercial software they used are accessible.

D. Capacity of generalization

Finally, we realized the first study on the generalization of DL methods for motion estimation in echocardiography. To this aim, we used two complementary datasets, the composition of which allowed us to conduct a multi-center, multi-vendor and multi-disease study. Tables IX to X illustrate the strong ability of our DL solution to provide accurate and consistent results for a wide range of situations, from different ultrasound machines to several pathologies with different motion patterns. This adaptability confirms the relevance of our transfer learning strategy as well as the quality of the synthetic data we have generated. To better assess the quality of the obtained results, several videos of tracking of the endocardial contours from both GE and Philips datasets are provided at: `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`

E. Perspectives

Although our results on the clinical data are convincing, we can observe a decrease of performance between simulation and

clinical data. There is therefore room for improvement. For a generalization point of view, it would be more interesting to have the simulated reverberation artifacts appear at random positions during training. Unfortunately, this implies the generation of new synthetic images when creating each new batch, which is currently not possible due to the computation time of the simulation (around 2 minutes per synthetic image). An intermediate possibility would be to simulate for each patient several sequences with reverberation artifacts at different positions and to draw the corresponding images randomly when creating the different batches. An alternative to enhance the generalization capability of DL methods would be to simulate a richer dataset by changing the motion or reflection of the scatterers to increase the variability of the image quality. It would also be interesting to integrate out-of-plane motions as they contribute to the deterioration of the motion estimation in 2D. In parallel, another way to increase the accuracy of our network would be to optimize more tasks on the same cine loop. For instance, joint optimization of motion estimation and segmentation tasks, as recently proposed in [22], may be an interesting track to investigate.

VI. Conclusion

In this paper, we developed a deep learning method for motion estimation in echocardiography. We showed that the combination of a customized version of PWC-Net with a new simulated synthetic dataset and a dedicated data augmentation strategy outperforms the current state-of-the-art methods, both for the tracking of endocardial borders and the estimation of the GLS and MAPSE indices. The genericity of our approach was also demonstrated from the first multi-center, multi-vendor and multi-disease study. The proposed synthetic dataset consists of 2D apical four chamber view sequences for 100 virtual patients with or without reverberation artifacts and with the corresponding myocardial displacement fields. For open science purposes, the full dataset can be directly accessed at `http://humanheart-project.creatis.insa-lyon.fr/medicaid.html`.

References

[1] V. Mor-Avi, R. M. Lang, L. P. Badano, M. Belohlavek, N. M. Cardim, G. Derumeaux, M. Galderisi, T. Marwick, S. F. Nagueh, P. P. Sengupta *et al.*, "Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: Ase/eae consensus statement on methodology and indications endorsed by the japanese society of echocardiography," *European Journal of Echocardiography*, vol. 12, no. 3, pp. 167–205, 2011.

[2] E. Evain, K. Faraz, T. Grenier, D. Garcia, M. De Craene, and O. Bernard, "A pilot study on convolutional neural networks for motion estimation from ultrasound images," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 12, pp. 2565–2573, 2020.

[3] J.-U. Voigt, G. Pedrizzetti, P. Lysyansky, T. H. Marwick, H. Houle, R. Baumann, S. Pedri, Y. Ito, Y. Abe, S. Metz *et al.*, "Definitions for a common standard for 2d speckle tracking echocardiography: consensus document of the eacvi/ase/industry task force to standardize deformation imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 1, pp. 1–11, 2015.

[4] K. E. Farsalinos, A. M. Daraban, S. Ünlü, J. D. Thomas, L. P. Badano, and J.-U. Voigt, "Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the eacvi/ase inter-vendor comparison study," *Journal of the American Society of Echocardiography*, vol. 28, no. 10, pp. 1171–1181, 2015.

[5] V. Perrot and D. Garcia, "Back to basics in ultrasound velocimetry: tracking speckles by using a standard piv algorithm," in *2018 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2018, pp. 206–212.

[6] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[7] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1879–1892, 2019.

[8] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[11] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4161–4170.

[12] D. Sun, X. Yang, M.-Y. Liu, and J. 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*, 2018, pp. 8934–8943.

[13] N. Duchateau, A. P. King, and M. De Craene, "Machine learning approaches for myocardial motion and deformation analysis," *Frontiers in Cardiovascular Medicine*, vol. 6, p. 190, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcvm.2019.00190

[14] H. Yu, X. Chen, H. Shi, T. Chen, T. S. Huang, and S. Sun, "Motion pyramid networks for accurate and efficient cardiac motion estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 436–446.

[15] M. Morales, M. van den Boomen, C. Nguyen, J. Kalpathy-Cramer, B. Rosen, C. Stultz, D. Izquierdo-Garcia, and C. Catana, "Deepstrain: A deep learning workflow for the automated characterization of cardiac mechanics," *bioRxiv*, pp. 2021–01, 2021.

[16] M. A. Morales, D. Izquierdo-Garcia, I. Aganj, J. Kalpathy-Cramer, B. R. Rosen, and C. Catana, "Implementation and validation of a three-dimensional cardiac motion estimation network," *Radiology: Artificial Intelligence*, vol. 1, no. 4, p. e180080, 2019.

[17] A. K. Z. Tehrani and H. Rivaz, "MPWC-Net++: evolution of optical flow pyramidal convolutional neural network for ultrasound elastography," in *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, B. C. Byram and N. V. Ruiter, Eds., vol. 11602. SPIE, 2021, pp. 14 – 23.

[18] M. G. Kibria and H. Rivaz, "Gluenet: Ultrasound elastography using convolutional neural network," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*. Springer, 2018, pp. 21–28.

[19] A. K. Z. Tehrani and H. Rivaz, "Displacement estimation in ultrasound elastography using pyramidal convolutional neural network," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2629–2639, 2020.

[20] B. Peng, Y. Xian, Q. Zhang, and J. Jiang, "Neural-network-based motion tracking for breast ultrasound strain elastography: An initial assessment of performance and feasibility," *Ultrasonic Imaging*, vol. 42, no. 2, pp. 74–91, 2020.

[21] S. S. Ahn, K. Ta, A. Lu, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, "Unsupervised motion tracking of left ventricle in echocardiography," in *Medical Imaging 2020: Ultrasonic Imaging and Tomography*, vol. 11319. International Society for Optics and Photonics, 2020, p. 113190Z.

[22] K. Ta, S. S. Ahn, A. Lu, J. C. Stendahl, A. J. Sinusas, and J. S. Duncan, "A semi-supervised joint learning approach to left ventricular segmentation and motion tracking in echocardiography," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1734–1737.

[23] A. Østvik, E. Smistad, T. Espeland, E. A. R. Berg, and L. Lovstakken, "Automatic myocardial strain imaging in echocardiography using deep learning," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 309–316.

[24] A. Østvik, I. M. Salte, E. Smistad, T. M. Nguyen, D. Melichova, H. Brunvand, K. Haugaa, T. Edvardsen, B. Grenne, and L. Lovstakken,

"Myocardial function imaging in echocardiography using deep learning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1340–1351, 2021.

[25] M. Alessandrini, B. Chakraborty, B. Heyde, O. Bernard, M. De Craene, M. Sermesant, and J. D'hooge, "Realistic vendor-specific synthetic ultrasound data for quality assurance of 2-d speckle tracking echocardiography: Simulation pipeline and open access database," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 65, no. 3, pp. 411–422, 2018.

[26] M. Alessandrini, M. De Craene, O. Bernard, S. Giffard-Roisin, P. Allain, I. Waechter-Stehle, J. Weese, E. Saloux, H. Delingette, M. Sermesant, and J. D'hooge, "A pipeline for the generation of realistic 3d synthetic echocardiographic sequences: Methodology and open-access database," *IEEE Transactions on Medical Imaging*, vol. 34, no. 7, pp. 1436–1451, 2015.

[27] S. Shahriari and D. Garcia, "Meshfree simulations of ultrasound vector flow imaging using smoothed particle hydrodynamics," *Physics in Medicine & Biology*, vol. 63, no. 20, p. 205011, oct 2018. [Online]. Available: https://doi.org/10.1088/1361-6560/aae3c3

[28] V. Perrot, M. Polichetti, F. Varray, and D. Garcia, "So you think you can das? a viewpoint on delay-and-sum beamforming," *Ultrasonics*, vol. 111, p. 106309, 2021.

[29] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, 2019.

[30] M. Alessandrini, A. Basarab, H. Liebgott, and O. Bernard, "Myocardial motion estimation from medical images using the monogenic signal," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1084–1095, 2013.

[31] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] H. Jacinto, R. Kéchichian, M. Desvignes, R. Prost, and S. Valette, "A web interface for 3d visualization and interactive segmentation of medical images," in *Proceedings of the 17th International Conference on 3D Web Technology*, 2012, pp. 51–58.

[33] D. Garcia, "A fast all-in-one method for automated post-processing of piv data," *Experiments in Fluids*, vol. 50, no. 5, pp. 1247–1259, May 2011.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[35] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[36] J. E. Sanderson and A. G. Fraser, "Systolic dysfunction in heart failure with a normal ejection fraction: Echo-doppler measurements," *Progress in Cardiovascular Diseases*, vol. 49, no. 3, pp. 196–206, 2006.

[37] K. Hu, D. Liu, S. Herrmann, M. Niemann, P. D. Gaudron, W. Voelker, G. Ertl, B. Bijnens, and F. Weidemann, "Clinical implication of mitral annular plane systolic excursion for patients with cardiovascular disease," *European Heart Journal–Cardiovascular Imaging*, vol. 14, no. 3, pp. 205–212, 2013.

[38] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Springer Berlin Heidelberg, 2003, pp. 363–370.