Myocardial Function Imaging in Echocardiography Using Deep Learning

Andreas Østvik, Ivar Mjåland Salte, Erik Smistad, Thuy Mi Nguyen, Daniela Melichova, Harald Brunvand, Kristina Haugaa, Thor Edvardsen, Bjørnar Grenne, and Lasse Lovstakken, *Member, IEEE*

***Abstract*— Deformation imaging in echocardiography has been shown to have better diagnostic and prognostic value than conventional anatomical measures such as ejection fraction. However, despite clinical availability and demonstrated efficacy, everyday clinical use remains limited at many hospitals. The reasons are complex, but practical robustness has been questioned, and a large inter-vendor variability has been demonstrated. In this work, we propose a novel deep learning based framework for motion estimation in echocardiography, and use this to fully automate myocardial function imaging. A motion estimator was developed based on a PWC-Net architecture, which achieved an average end point error of (0.06 ± 0.04) mm per frame using simulated data from an open access database, on par or better compared to previously reported state of the art. We further demonstrate unique adaptability to image artifacts such as signal dropouts, made possible using trained models that incorporate relevant image augmentations. Further, a fully automatic pipeline consisting of cardiac view classification, event detection, myocardial segmentation and motion estimation was developed and used to estimate left ventricular longitudinal strain in vivo. The method showed promise by achieving a mean deviation of $(-0.7 \pm 1.6)\%$ compared to a semi-automatic commercial solution for N = 30 patients with relevant disease, within the expected limits of agreement. We thus believe that learning-based motion estimation can facilitate extended use of strain imaging in clinical practice.**

***Index Terms*— Deep learning, echocardiography, motion estimation, deformation, strain.**

Manuscript received December 1, 2020; revised January 13, 2021; accepted January 20, 2021. Date of publication January 25, 2021; date of current version April 30, 2021. This work was supported in part by the Research Council of Norway under Project 237887 and in part by the Norwegian Health Association, South-Eastern Norway Regional Health Authority and the National Programme for Clinical Therapy Research (KLINBEFORSK) under Project 2017207. *(Corresponding author: Andreas Østvik.)*

Andreas Østvik and Erik Smistad are with the Centre for Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, 7491 Trondheim, Norway, and also with SINTEF Digital, Department of Health Research, 7465 Trondheim, Norway (e-mail: andreas.ostvik@ntnu.no; erik.smistad@ntnu.no).

Ivar Mjåland Salte, Thuy Mi Nguyen, and Daniela Melichova are with the Department of Medicine, Sørlandet Hospital Kristiansand, 4615 Kristiansand, Norway, and also with the Faculty of Medicine, University of Oslo, 0372 Oslo, Norway (e-mail: ivar.mjaland.salte@sshf.no; thuy.mi.nguyen@sshf.no; daniela.melichova@sshf.no).

Harald Brunvand is with the Department of Medicine, Sørlandet Hospital Kristiansand, 4615 Kristiansand, Norway (e-mail: harbrun@online.no).

Kristina Haugaa and Thor Edvardsen are with the Department of Cardiology, Oslo University Hospital, Rikshospitalet, 0372 Oslo, Norway, and also with Faculty of Medicine, University of Oslo, 0372 Oslo, Norway (e-mail: thor.edvardsen@medisin.uio.no; kristina.haugaa@medisin.uio.no).

Bjørnar Grenne is with the Centre for Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, 7491 Trondheim, Norway, and also with the Clinic of Cardiology, St. Olavs Hospital, 7030 Trondheim, Norway (e-mail: bjornar.grenne@ntnu.no).

Lasse Lovstakken is with the Centre for Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: lasse.lovstakken@ntnu.no).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3054566, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3054566

I. Introduction

MOTION estimation is an essential part of ultrasound imaging, especially in echocardiography, where it is used to assess cardiac function. Currently speckle tracking echocardiography (STE) is widely deployed, with many methodological variants such as variational optical flow (OF) and block-matching methods [1]. In research, conventional STE methods have been outperformed by phase sensitivity and elastic registration methods [2]–[4]. Despite being considered the standard, these methods have several unsolved challenges due to fundamental limitations of ultrasound (US) acquisitions. This includes dropouts, shadows, out-of-plane motion, drift sensitivity, foreshortening and more [5]. Several of these artifacts leads to a decorrelation of the US speckle pattern from frame to frame, thus complicating the tracking task.

Deformation imaging, such as measurement of myocardial strain, has shown great potential [6]–[8], and is claimed to have better diagnostic and prognostic value compared to conventional anatomical measurements such as ejection fraction (EF). Motion estimation (ME) is usually an essential part of these methods, and the measurements are dependent on its performance. Clinical use of deformation imaging is still limited, partly due to time constraints in the clinic, but also a lack of consensus about robustness and reproducibility. We also hypothesize that the retrospective nature of the analysis reduces its use, and believe that having the possibility to quality assure acquisitions while scanning would facilitate clinical implementation. Major efforts have been

put into standardization of strain estimation techniques [8], [9]. Part of this involves developing common evaluation platforms and data, in which Alessandrini *et al*. [10] proposed a realistic *in silico* database of US sequences based on simulations with biomechanical models for comparison of STE algorithms.

Recently, motion estimation using convolutional neural networks (CNN) have shown promising results for general optical flow (OF) problems. Dosovitskiy *et al*. [11] demonstrated this, by learning to estimate motion patterns directly from images using U-Net based architectures called FlowNet. Several flavors of the topology exists, such as FlowNetS, FlowNetC and FlowNet-SD, with decisive modifications, for instance to resolve issues with noisy artifacts and small displacements. By stacking several of these networks in a cascade and using complex training schedules, as in FlowNet 2.0, performance was on par or better than state of the art methods for traditional OF estimation. These methods introduced a shift in OF research, and in few years the work on the topic has increased dramatically, where the benchmarks have been dominated by deep learning (DL) based methods. One of the limitations of FlowNet 2.0 is the network complexity and inference speed. In PWC-Net, the developers succeeded in both increasing the accuracy and reducing the size of the CNN model by leveraging conventional OF components [12]. The PWC-Net and FlowNet architectures are currently the most common starting point for research on DL based OF estimation. The main difference between them is that FlowNet is encoder-decoder based, while PWC-Net use a spatial pyramid.

Using these type of network designs directly for ME in US imaging raises some concerns. Firstly, their design and training regime facilitate correlation between global image features, and an optimization for rigid motion patterns. This is not fully compatible with deformation imaging, where local coherent speckle is used to track local tissue motion and inherent non-rigid deformation patterns. Structures in US images do not have clear borders and traditionally STE has relied on tracking the local speckle pattern, rather than global texture features. On the other hand, speckle decorrelation occurs throughout the cardiac cycle, and this is a fundamental limit of static tracking kernels. Current CNN methods for ME use block matching between features of consecutive frames, but not between lower levels of the architecture [12], [13]. This does not distinguish noise and speckle locally, and the cost volume will thus make limited use of coherent speckle between consecutive frames. We thus hypothesize that learning-based ME extended with knowledge from STE methods could improve robustness, and therefore be beneficial.

The use of deep learning based ME in US, and especially in echocardiography, is limited [14]. Earlier, we demonstrated the use of FlowNet 2.0 out-of-the-box for estimating global longitudinal strain (GLS) in a pipeline with view classification, segmentation of the myocardium and state-estimation techniques [15]. The results were promising, but both the training data and methods had several limitations, especially for regional motion patterns. In elastography, several studies have been conducted with use of FlowNet 2.0 to estimate the displacements [16]. In a recent pilot study, efforts were also made into benchmarking different networks components of FlowNet 2.0, with fine-tuning on simulated US data. The results were on par with current state of the art for flow estimation [17]. In sum, these studies indicate a potential and adaptability for CNN based ME in US image analysis.

Utilizing DL models in a cascade for fully automating clinical measurements have also become a popular research topic, for instance for measurements such as EF and strain [18]. We recently demonstrated an accuracy within interobserver variability on calculations of EF, with possibility for real-time analysis and quality assurance on-site [19]. In this study we aim to extend our work by incorporating ME in an automatic pipeline, in order to do fully automatic deformation measurements. Our goal is to develop DL based methods which may facilitate the implementation of functional imaging in the clinic by removing several steps of manual post-processing and enable real-time use. This could make the measurements more robust and less time consuming.

A. Main Contributions

We propose a novel framework for motion estimation in echocardiography, and use this, together with other relevant components, to fully automate the estimation of longitudinal strain. The contributions of this paper are

- A motion estimator for echocardiography inspired by PWC-Net that incorporates domain knowledge from US, and constraints from relevant morphophysiology.
- A training setup with pretraining on synthetic data, and finetuning on more realistic US simulations with relevant augmentation routines.
- Analysis of deep learning based motion estimation with comparison on simulated and in-vivo data.
- A fully automated pipeline for longitudinal strain measurements using cardiac view classification, event detection, segmentation and motion estimation by DL.
- Comparison between the automated pipeline and a commercial available system for GLS measurements.

II. Methods

A. Motion Estimation With Deep Learning

Currently, PWC-Net is the most popular architecture for deep learning based optical flow estimation, and several variations exist [12], [20], [21]. It is inspired by conventional OF, including components such as pyramidal coarse-to-fine estimators, warping and cost volume in the design pattern. Still, it utilizes the strengths of CNNs by incorporating feature learning in several stages.

A simplified illustration of the network architecture can be seen in the left part of Fig. 1. The core method involves taking two consecutive images as input, and these are fed separately into a learnable CNN based feature extractor pyramid of L levels with shared weights. At each level l, the feature maps from the previous level is downsampled to half its size using strided convolutions. The features of level l of the second image is warped towards the first image using the upsampled flow from the consecutive level. A cost volume is estimated

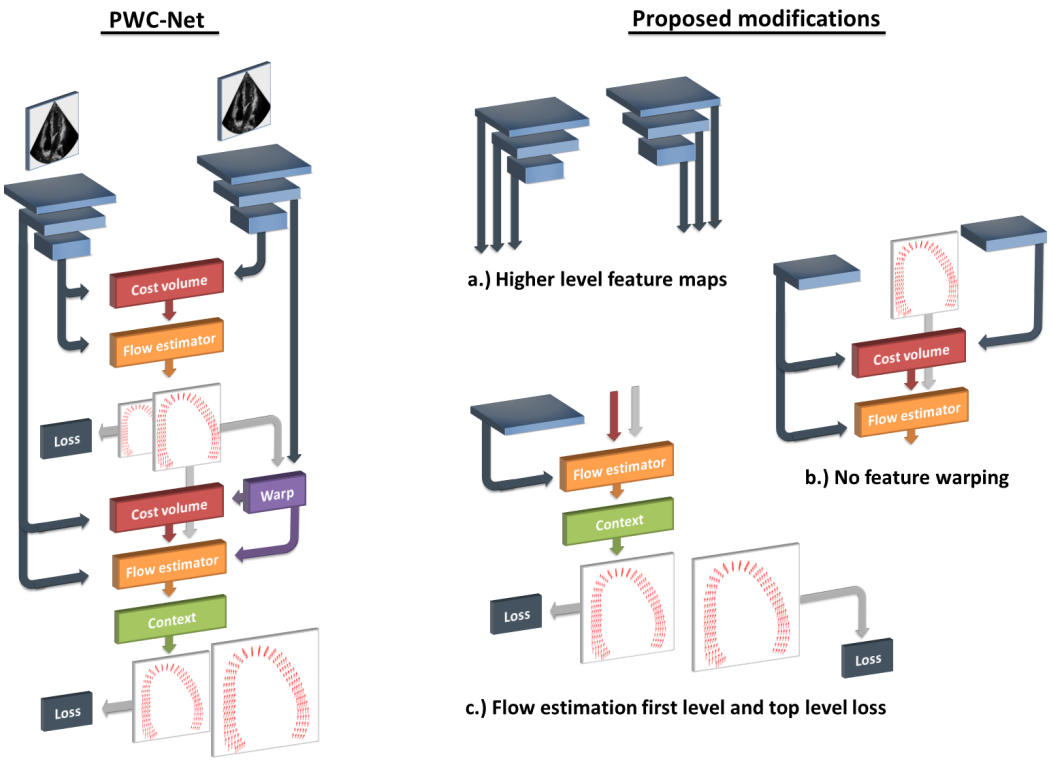


Fig. 1. Sketch of a traditional PWC-Net architecture with three pyramid levels (left). Two consecutive images are fed into a pyramidal feature extractor. The cost volume is estimated between feature maps of the first image and the backward warped second image features (no warping at bottom). A CNN named Flow estimator, is used to estimated the flow at every level. At the top level a context network is used to refine the flow. The right part of the figure illustrates the modifications done for the EchoPWC-Net. Firstly, feature maps closer to the input is propagated through the cost volume and flow estimation routines (a). Warping of the features of the second image is removed and exchanged with a direct correlation between features (b) and flow at the two highest levels is also included in the loss (c).

using correlation between the first image and warped features of the second image. For each layer, the cost volume, features from the first image and upsampled optical flow are input into a CNN which outputs a dense displacement map for the current pyramid level. The estimation is repeated upwards in size until the desired level. The output is then forwarded into a context network with dilated convolutions, which refines the flow, taking the estimated flow and features of the second last layer from the OF estimator as input. The final output is a dense displacement map resized to the same spatial size as the input images.

The original PWC-Net implementation has seven feature pyramid extractor levels including the inputs. The output level is one-quarter of the inputs spatial size, and the flow is upsampled by bilinear interpolation after the context network. The basis of our implementation also has seven pyramid levels, but as opposed to the original implementation which produces flow estimation up to the second highest level, we extend our network to produce flow estimation up to the first level. This is further fed into a context network before the final upsampling as indicated in the right side of Fig. 1. Also, we include the final output in the loss function. This is to retain some of the useful speckle patterns lost when resampling from a low resolution level, and optimize for local variations and small displacements. We also hypothesise that the ambiguity caused by occlusions and out-of-plane motion during warping makes the original implementation problematic for echocardiography. One of the main motivations of using warping is to handle large motions and allow for smaller networks, but for echocardiography the typical motion between frames is small. We therefore remove the warping procedure, and instead estimate the cost volume directly between feature maps at every level. This is illustrated in Fig. 1(b). We argue that this integrates some of the benefits of block matching between frames at every level of the pyramid, and thus have more resemblance to traditional STE. However, instead of locating the minima of the cost volume as you would with STE, the full correlation map is passed to the flow estimator. We refer to our customized network as EchoPWC-Net. Additional implementation details are given in Section III-C.

B. Pipeline for Automated Functional Imaging

The proposed pipeline for myocardial function imaging is summarized in Fig. 2. Together with the discussed motion estimation, it consists of several in-house DL based methods, including cardiac view classification, event detection and myocardial segmentation. In addition, we initialize the tracking by extracting the mid ventricular centerline of the myocardium. We summarize the steps in the following, and the reader is referred to published work for more details about the DL networks [22]–[24].

1) View Classification: To ensure valid acoustic windows, we employ an in-house cardiac view classification (CVC)

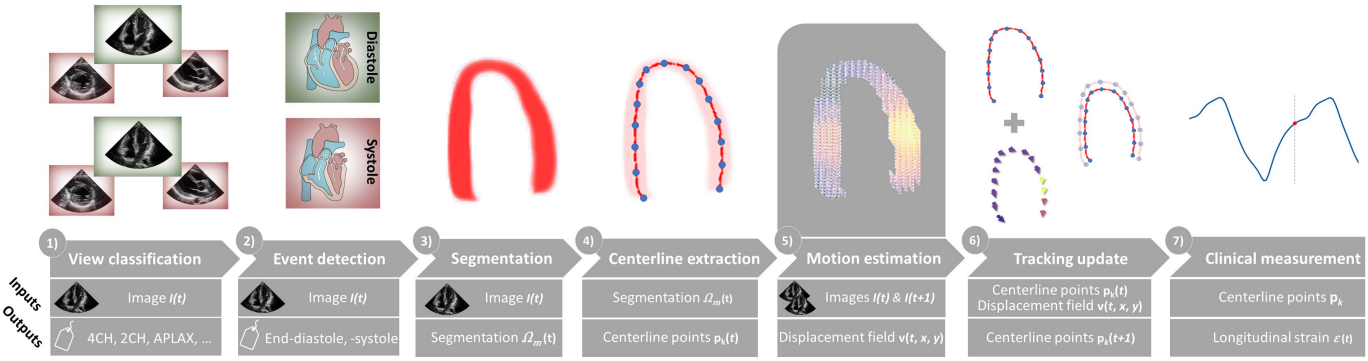


Fig. 2. The measurement pipeline. Valid US images are forwarded through a segmentation network, and the resulting masks are used to extract the centerline and relevant parts of the image. The US data is further processed through the motion estimation network yielding a map of velocity vectors. The centerline is used to seed points which are used for tracking the myocard. The velocities of the myocard are optionally used either directly to propagate the centerline points, or as a measurement update step of a Kalman filter. The results are used as a basis for strain measurements.

network [22]. The method recognizes up to eight different cardiac views, including the apical four-chamber (4CH), apical two-chamber (2CH), apical long-axis (APLAX), which are relevant for this study. The network topology is composed of seven block levels of convolution filters, batch normalization, PReLU activation and max pooling. Inception modules and a dense connectivity pattern are employed in the last five blocks. A global average pooling layer was used before the final softmax activation. It was trained on a dataset of approximately 250 patients, and tested on a similarly sized independent dataset. The network input is standard scan converted B-mode images of size (128×128), and the output is a softmax activation yielding a confidence score for each class. This network has shown an accuracy of 98% and inference time of approximately 4 ms per frame.

2) Event Detection: The cardiac phases are identified using a sequence-to-sequence CNN that can classify diastole and systole directly from B-mode images [23]. The network consists of five stacked levels of 3D convolutions, batch normalization, ReLU activation and max pooling. All convolution kernels have a temporal size of three, while the spatial kernel size is (7×7) for the first layer and (3×3) for the rest. The output of this stage is then propagated into two layers of long short-term memory (LSTM) modules with 32 units each. It was trained and validated on the CAMUS dataset of 500 patients [25]. The network handles variable number of frames with size (128×80) as input, and outputs a sequence of scalars in the interval zero to one. Zero indicates that the image is from the systolic phase, while one is the diastolic phase. End-diastolic (ED) and end-systolic (ES) frames were identified as the temporal points where the phase changes, i.e. cross-over from zero to one and vice versa. The method has shown an accuracy of (-5.5 ± 28.2) ms and (-0.6 ± 31.8) ms on ED and ES frames respectively, and mean absolute error of 1.53 and 1.55 frames from reference. For batch processing, a runtime of 16 ms per frame was measured.

3) Myocard Segmentation: We utilize a segmentation network proven to work well in several studies. This is a slight modification of the U-Net architecture [26], with six levels in the encoder and decoder part. Each level is composed of (3×3) convolution filters and ReLU activation. Max pooling is performed in the encoder part, while upsampling with nearest neighbor interpolation is performed in the decoder part. Skip connections are used between the levels at each stage. The network was first described by Smistad *et al.* [27] and later used in the CAMUS study of Leclerc *et al.* [25]. Recently, it has been used with success in an automatic measurement pipeline for ejection fraction and foreshortening detection [19]. In this study, we use the segmentation of the myocardium Ω_m. Initially, the network was trained for 4CH and 2CH views, but it was later extended to include the APLAX view [24]. It was designed for real-time performance, with 2 million parameters. Network input is an US image of size (256×256) together with a binary value indicating if it is an APLAX view or not. The output is a map of same size of the input image, where each pixel is classified as either LV lumen, myocard, left atrium or background. Data from the CAMUS dataset of 500 patients together with parts of an internal study were used for training. The network achieved a test dice score of 0.79 on the myocardium. A runtime of about 10 ms on a GPU was achieved.

4) Centerline Extraction: The centerline $\mathcal{C}$ of the myocardium is defined by extracting the contour of the myocardial segmentation Ω_m and defining the endo- and epicardial borders. Further the base and apex points are defined as the points furthest away from the LV lumen centroid, in left bottom, right bottom and top direction respectively. The centerline is defined as the mid-point between two nearest endo- and epicardial points on the line perpendicular to the longitudinal. A total of k equidistant points $\mathbf{p}_k = \langle x, y \rangle$ along the longitudinal direction is then sampled, i.e. $\mathcal{C} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k\}$.

5) Motion Estimation: The pipeline allows using different motion estimation methods. In this study, we employ four different variants, a traditional Farnebäck optical flow method [28], the FlowNet 2.0, the original PWC-Net and a modified PWCNet which we named EchoPWC-Net. With Farnebäck we use a grid based optimization minimizing average end point error (EPE) on simulated data to find the parameters for window size, pyramid levels, pyramid scale, iterations at pyramid scale, size of the kernel for polynomial expansion and smoothing factor for the derivative of the

polynomial. All the methods produce a dense displacement map of velocity components $\mathbf{v}(x, y) = \langle v_x, v_y \rangle$ between two images $I(t)$ and $I(t+1)$.

6) Tracking Update: The centerline points $\mathbf{p}_k$ can either be updated by propagating the points with the displacement field, i.e. $\mathbf{p}_k(t+1) = \mathbf{p}_k(t) + \mathbf{v}_k(t)$. Alternatively, it can be extracted from the segmentation directly without using the motion estimation method at all.

This step could also involve state estimation techniques such as the Kalman filter [29] or similar, but this was not pursued further in this study.

7) Clinical Measurements: The centerline $\mathcal{C} \subseteq \Omega$ is used to calculate the longitudinal ventricular length ι, i.e. the arc length, for each timestep t. Further, this is used to estimate the Lagrangian strain

$$\epsilon(t) = (\iota(t) - \iota_0)/\iota_0, \tag{1}$$

along the center of the myocard. The reference length ι_0 is measured at the ED frame. The peak-systolic strain was used for both GLS and regional longitudinal strain (RLS) estimation, where the peak was defined as the minima between ED and ES strain values. For RLS, we divide the ED centerline at the apex and estimate three equally sized arcs on both sides and compute their strain individually [9].

III. Experiments

A. Datasets

Several datasets were used developing the methods, and in the following we will briefly describe the datasets used for modeling the motion estimation network, and testing the measurement pipeline. For information about data used to train other models, such as segmentation, view classification and event detection, the reader is referred to publications on the specific networks [22]–[24].

1) Synthetic Data: We used three publicly available datasets commonly used for training and benchmarking of optical flow methods. All the datasets consists of image pairs and a corresponding dense displacement map.

- *FlyingChairs2D [11]:* Contains images of rendered 3D chair models moving in front of random backgrounds scraped from the photo management and sharing site Flickr. A total of 22872 images.
- *FlyingThings3D [30]:* Contains approximately 25000 stereo images sampled from a 3D scene of everyday objects flying along randomized trajectories on a textured background.
- *MPI SINTEL [31]:* Contains images from an open source animated short film. A total of 1628 frames from 35 different animation scenes.

Example image pairs from the datasets with corresponding flow can be seen in the upper part of Fig. 3.

2) Simulated Ultrasound Data: An open database of simulated echocardiography images created for quality assurance of speckle tracking algorithms [10] were employed. The data is created with a complex simulation pipeline, where a 3D dataset of simulated US volumes of the heart

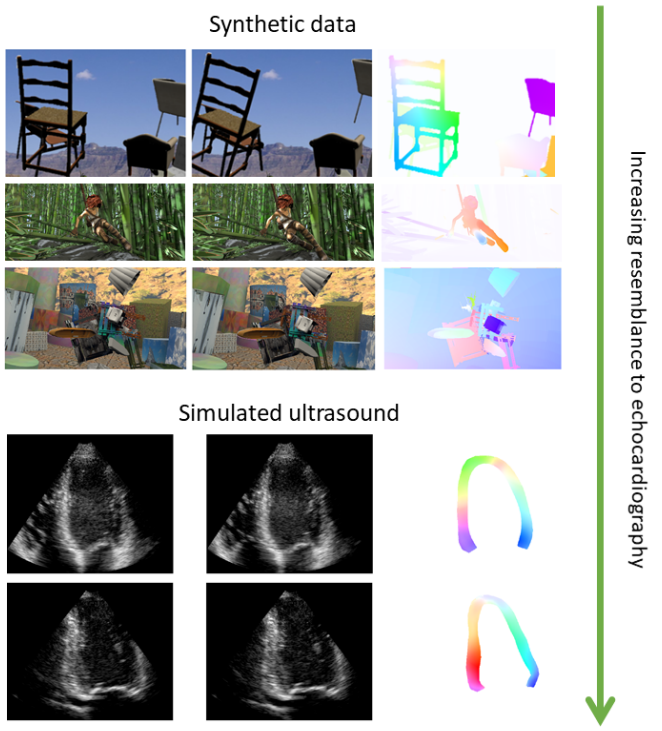


Fig. 3. Examples from the datasets. Synthetic data from FlyingChairs2D [11], FlyingThings3D [30] and MPI SINTEL [31] and simulated ultrasound [10]. For each example, from left to right, we have two consecutive frames followed by the flow field from the first to the second frame.

and corresponding myocardial mesh is spatio-temporally aligned with a 2D template of real US data. Further, a synthetic motion field from a biomechanical model was used to propagate the mesh and aligned data. A scatter map was generated from the composition, and used to generate simulated US. In total, the data is composed of templates from seven different vendors and five motion patterns from the biomechanical model, including one healthy and four pathologies. Each with the three apical views, 4CH, 2CH and APLAX, resulting in a total of 105 sequences or 6165 frames. For each timestep, a set of 180 points divided among five longitudinal lines and six segments is provided by the authors. These points correspond to the underlying motion field of the biomechanical model aligned with the US data. An example of image pairs of 4CH and 2CH views from the dataset with corresponding flow can be seen in the lower part of Fig. 3. They also provide the view and the cardiac event timing for each sequence.

3) Clinical Data: A dataset was collected from a clinical database of patients diagnosed with acute myocardial infarction (MI) or de-novo heart failure (HF) at a Norwegian hospital. The study was approved by the regional ethics committee (ref. 2013/573) and written consent was given by all patients. The images were acquired using GE Vingmed (Vivid 7, E9 or E95) scanners. Patients were included consecutively regardless of image quality. All exams were performed in clinically stable patients with sinus rhythm. Images were analyzed by a single clinician using clinical best practice as defined in [7] with the 2D strain (2DS) application in the clinical software

EchoPAC release 202 (GE Vingmed AS, Horten, Norway.). This ensures that the proposed automated method is compared to actual clinical practice measurements techniques. To ensure a representative range of LV pathologies, a total of 30 patients from five different cohorts were randomly selected, resulting in six patients from each group. The groups were defined by a diagnosis of ST elevation MI (STEMI), non-ST elevation MI (NSTEMI), ischemic heart failure (HF), non-ischemic HF and no significant disease. The tracked mid ventricular points and corresponding strain values were exported from the software.

B. Data Augmentation

Due to the unrealistic nature of simulated ultrasound and limited access to relevant (in vivo) data with a ground truth, we rely on several US-specific augmentation routines. Here we try to induce realistic artifacts common in echocardiography that usually hampers the success of speckle tracking, and describe a selection in the following.

1) Gaussian Shadowing: Acoustic shadows often occur in US imaging due to structures that strongly reflect or absorb the US waves. This is often identified as a dark region behind the structure. We mimic this effect by placing random regions of intensity reductions in the image. Similar methods have been shown to have an effect on generalization for US segmentation tasks [32].

2) Haze Artifact Application: One artifact that is prevalent for some patient is acoustic haze. This can be identified as a semi-static noise band in the upper parts of the image. We randomly apply static high intensity artifacts with a Gaussian profile along the radial direction in polar coordinates.

3) Depth Attenuation: The US wave looses energy as it travels through the body, and this can be identified as a gradual drop in intensity with distance from the probe. Similarly like the haze artifact application, we apply a varying degree of intensity attenuation along the radial direction. The attenuation does not consider depth independent noise, and is thus a simplification of the physical artifact.

4) Speckle Reduction: The speckle pattern in images from different vendors often differ due to image enhancement and various filtering methods. To reproduce this effect, we smooth the images randomly using a bilateral filter, effectively reducing the speckle.

In addition to these US specific augmentations, we apply basic augmentations such as horizontal and vertical flipping, temporal reversing, frame skipping, rotation, random noise, scaling, image resampling artifacts, JPEG compression and gamma intensity transformations. Except for the flipping, reversing, scaling, skipping and rotation, the displacement map was not modified for any of the augmentation routines. All augmentations are applied in random combinations and on-line while training. Examples from some of the individual augmentations are given in the supplementary material. In addition, the effect on maximum flow distribution after five epochs with augmentation is visualized.

C. Implementation Details

We implemented the machine learning environment using Tensorflow [33] version 2. The modeling and experiments were conducted on a workstation with an Ubuntu 16.04 operating system. The hardware consisted of an Intel Xeon CPU E5-2637 v2 with a clock speed of 3.50 GHz, 112 GB RAM and a NVIDIA Titan V GPU with 12 GB of memory.

1) Architecture Parameters: For the feature extractor in EchoPWC-Net, we use one convolution layer in addition to the strided convolution, with equal amount of filters at each level. The amount of filters used was 16, 32, 64, 96, 128 and 192, from top to bottom level respectively. For the cost volume we use a search range of 4 for every level, and for the context network we use the architecture proposed in the original implementation [34]. This corresponds to a receptive field of 67×67 in the last layer.

2) Data Preprocessing: To reduce the feature space and adapt for US, we converted all input data to grayscale, including the synthetic RGB data. The input was set to a fixed size of (448×576). As mentioned, the simulated US data is provided together with a set of 180 spatial points inside the myocardium for every timestep. We use these to generate a sparse displacement field, and we use cubic interpolation to convert to a dense displacement map with velocities $\mathbf{v}(x, y) = \langle v_x, v_y \rangle$ inside the myocardium Ω_{m}. To avoid boundary effects, we extrapolate the epicardial points radially by 5% of the radial diameter, followed by masking by the concave hull enclosing the original points. The dense displacement map is used as the ground truth flow, and the units were set to pixel per frame.

3) Training Procedures: Several training dataset schedules were investigated, resulting in four different setups:

- **Synthetic RGB:** Sequential training from scratch with FlyingChairs2D and FlyingThings3D RGB data.
- **Synthetic gray:** Sequential training from scratch with grayscale FlyingChairs2D and FlyingThings3D.
- **Synthetic gray → Simulated US:** Initialized with weights from synthetic gray followed by fine-tuning on simulated US.
- **Simulated US:** Trained from scratch on simulated US.

When training with synthetic data, we employ the basic augmentations mentioned in Section III-B. In addition, we employ the US specific augmentation when training on simulated US.

Our models are trained with the Adam optimizer and a batch size of 4 for all experiments. The initial learning rate was set to 10^{-4}, with a halving schedule for each 100k and 20k iterations for synthetic and simulated US respectively. For fine-tuning, the initial learning rate was set to 10^{-5}. Training time from scratch was approximately three-four days for synthetic data with the fine-tuning schedule running over five days, and two days for simulated US. Early stopping with a patience of 30 epochs were used for all models.

4) Loss: We use a multi-scale loss function with end-point error. Since the labeled motion is sparse, we only optimize regionally where the input lies within the predefined segmented region Ω_{m}. We let $\mathbf{w}^l$ denote the dense flow field at

TABLE I
OVERVIEW OF DIFFERENT MOTION ESTIMATION MODELS

	Training dataset schedule			Augmentation	
	Synth. RGB	Synth. gray	Sim. US	Basic	US spec.
Farnebäck					
FlowNet 2.0	✓			✓	
PWC-Net	✓			✓	
PWC-Net-gray		✓		✓	
PWC-Net-gray-usft		✓	✓	✓	✓
PWC-Net-us			✓	✓	✓
EchoPWC-Net		✓		✓	
EchoPWC-Net-usft		✓	✓	✓	✓
EchoPWC-Net-us			✓	✓	✓

the lth pyramid level. The loss is defined as

$$\mathcal{L}(\Theta) = \sum_{l=l_0}^{L} \beta_l \sum_{\mathbf{x}} |\tilde{\mathbf{w}}^l_\Theta(\mathbf{x}) - \mathbf{w}^l_{GT}| + \gamma\,|\Theta| \quad \forall \quad \mathbf{w} \in \Omega_{\mathrm{m}},$$

where Θ is the parameters and $\mathbf{x}$ is the inputs. The term β_l is set manually and used to weight the loss contribution from each layer. The second term regularizes the parameters, where γ is the regularization factor. This is similar to the loss used in FlowNet and PWC-Net, but restricted to regional optimization. In our implementation, we set the weights to $\beta_0 = 0.015$, $\beta_1 = 0.03$, $\beta_2 = 0.06$, $\beta_3 = 0.12$, $\beta_4 = 0.25$, $\beta_5 = 0.50$ and $\beta_6 = 1.0$. Based on the input size, this correspond to equally weighting each layers contribution to the loss. The regularization factor γ was set to 10^{-4}.

D. Evaluation

1) Metrics: We evaluate our methods using the end point error (EPE), which is a common metric for benchmarking optical flow performance. It is defined as the Euclidean distance between the ground truth velocity and the predictions, i.e. EPE $= ||\mathbf{v}_{\mathrm{GT}} - \mathbf{v}_{\mathrm{pred}}||$. We also compute the strain values, as defined in (1). Regional strain is computed for each segment for each view, while global strain is computed for each view, and averaged over all views. In addition, we report correlation metrics, such as regression slope α and correlation coefficient ρ, as well as bias μ and 95 percentile limits of agreement (LOA).

2) Comparison I: Motion Estimation: Nine different motion estimation methods are evaluated. The original PWC-Net, as well as different flavours of the EchoPWC-Net. For reference, the Farnebäck and FlowNet 2.0 methods are also included. The various methods are summarized in Table I.

3) Comparison II: Automatic Pipeline: For functional measurements on in vivo data, we also test two variants of the presented pipeline:

- **Segmentation only**: Recalculation of the centerline for every time point, and not using motion estimation. This refers to skipping part 5) of the measurement pipeline.
- **Tracking**: Initialization of centerline by segmentation, and propagation of points using the best performing motion estimation model. This refers to the full pipeline described earlier.

4) Comparison III: Model Adaption: As mentioned, one of the limitations of traditional speckle tracking is the adaptability to various noise prevalent in US. To study the investigated methods ability to regularize, we design an evaluation strategy based on three of our US relevant augmentation routines, namely Gaussian shadow, haze artifact application and depth attenuation. More specifically, for Gaussian shadowing we apply a shadow region at the center of the mid septal segment of test data samples, and measure the change in relative EPE as a function of shadow amplitude, i.e. the relative degree of intensity signal. The size of the region was set to 20% of the image size in both directions, tuned to cover the whole segment for higher shadow amplitudes. For haze, we apply a localized band of haze mimicking noise in the upper half of the sector with an increasing intensity value. Finally, for depth attenuation we attenuate the intensity values gradually, with a fixed saturation area at the base level of the myocardium.

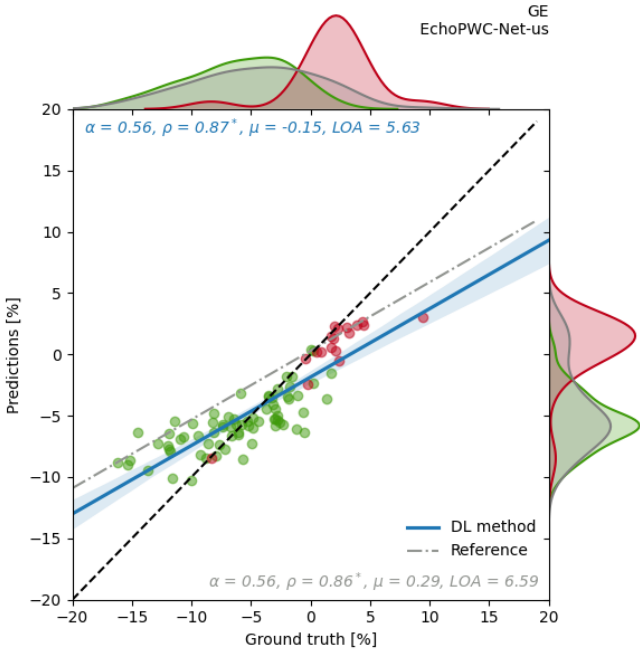


Fig. 4. Correlation plot between the ground truth regional strain estimation and the DL method on simulated data from one selected vendor. Green dots represent healthy myocardial segments, while the red sick segments. In the top left corner, the slope α of the regression line, correlation coefficient ρ, bias μ and limits of agreement (LOA) is given. The corresponding reference values from Alessandrini *et al.* [10] are given in the bottom right corner.

IV. RESULTS

A. Simulated Ultrasound

The PWC-Net model trained on grayscale FlyingChairs and FlyingThings3D achieved an average EPE of 4.80 and 6.41 on MPI Sintel Clean and Final respectively. Cross-validation was performed on the simulated US data by dividing into folds by vendor. This resulted in seven training sessions for each DL method. For the Farnebäck method, the grid based optimization yielded best EPE for 3 pyramid levels, with a scale of 0.5 and a window size of 69. A total of 5 iterations for each scale, a size of 5 pixels of the kernel for polynomial expansion and a smoothing factor of 1.1. The average EPE with corresponding standard deviation can be seen in Table IIa. On the respective test data, the results for segments and views

TABLE II
RESULTS ON SIMULATED ULTRASOUND DATA. AVERAGE END POINT ERROR (EPE) FOR (A) EVERY VENDOR, (B) AVERAGE OVER SEGMENTS AND (C) APICAL VIEWS. UNITS GIVEN IN MM PER TIMESTEP/FRAME ΔT^{-1}

(a) Vendors

Method	ESAOTE [mm·ΔT^{-1}]	GE [mm·ΔT^{-1}]	Hitachi [mm·ΔT^{-1}]	Philips [mm·ΔT^{-1}]	Siemens [mm·ΔT^{-1}]	Toshiba [mm·ΔT^{-1}]	Samsung [mm·ΔT^{-1}]
Farnebäck	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)
FlowNet 2.0	0.12 (0.10)	0.17 (0.13)	0.10 (0.08)	0.11 (0.08)	0.09 (0.08)	0.10 (0.08)	0.11 (0.09)
PWC-Net	0.12 (0.09)	0.13 (0.10)	0.10 (0.07)	0.10 (0.07)	0.09 (0.07)	0.10 (0.07)	0.10 (0.08)
PWC-Net-gray	0.19 (0.16)	0.21 (0.19)	0.13 (0.09)	0.15 (0.10)	0.12 (0.09)	0.15 (0.12)	0.17 (0.13)
PWC-Net-gray-usft	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)
PWC-Net-us	0.10 (0.08)	0.12 (0.10)	0.10 (0.08)	0.10 (0.08)	0.10 (0.08)	0.09 (0.07)	0.09 (0.07)
EchoPWC-Net	0.17 (0.17)	0.19 (0.20)	0.12 (0.09)	0.14 (0.11)	0.12 (0.09)	0.13 (0.10)	0.13 (0.10)
EchoPWC-Net-usft	0.09 (0.11)	0.11 (0.12)	0.09 (0.08)	0.09 (0.08)	0.10 (0.08)	0.08 (0.06)	0.07 (0.07)
EchoPWC-Net-us	**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)**

(b) Segments

Method	Base [mm·ΔT^{-1}]	Mid [mm·ΔT^{-1}]	Apical [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
Farnebäck	0.10 (0.07)	0.07 (0.05)	0.05 (0.04)	0.07 (0.06)
FlowNet 2.0	0.15 (0.10)	0.12 (0.08)	0.07 (0.07)	0.12 (0.09)
PWC-Net	0.14 (0.09)	0.10 (0.07)	0.08 (0.07)	0.11 (0.07)
PWC-Net-gray	0.21 (0.16)	0.14 (0.11)	0.12 (0.11)	0.16 (0.13)
PWC-Net-gray-usft	0.19 (0.12)	0.15 (0.10)	0.09 (0.07)	0.14 (0.10)
PWC-Net-us	0.14 (0.09)	0.10 (0.07)	0.06 (0.05)	0.10 (0.08)
EchoPWC-Net	0.19 (0.16)	0.13 (0.10)	0.10 (0.08)	0.14 (0.11)
EchoPWC-Net-usft	0.12 (0.11)	0.09 (0.08)	0.08 (0.07)	0.10 (0.08)
EchoPWC-Net-us	**0.08 (0.06)**	**0.06 (0.04)**	**0.04 (0.03)**	**0.06 (0.04)**

(c) Views

Method	4CH [mm·ΔT^{-1}]	2CH [mm·ΔT^{-1}]	APLAX [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
Farnebäck	0.07 (0.05)	0.08 (0.06)	0.08 (0.06)	0.07 (0.06)
FlowNet 2.0	0.10 (0.08)	0.12 (0.10)	0.12 (0.09)	0.12 (0.09)
PWC-Net	0.10 (0.07)	0.11 (0.08)	0.11 (0.08)	0.11 (0.08)
PWC-Net-gray	0.15 (0.11)	0.17 (0.15)	0.16 (0.12)	0.16 (0.13)
PWC-Net-gray-usft	0.13 (0.09)	0.14 (0.10)	0.15 (0.10)	0.14 (0.10)
PWC-Net-us	0.10 (0.07)	0.10 (0.08)	0.10 (0.08)	0.10 (0.08)
EchoPWC-Net	0.13 (0.12)	0.14 (0.16)	0.14 (0.14)	0.14 (0.14)
EchoPWC-Net-usft	0.11 (0.08)	0.10 (0.09)	0.10 (0.09)	0.10 (0.09)
EchoPWC-Net-us	**0.06 (0.04)**	**0.06 (0.04)**	**0.06 (0.05)**	**0.06 (0.04)**

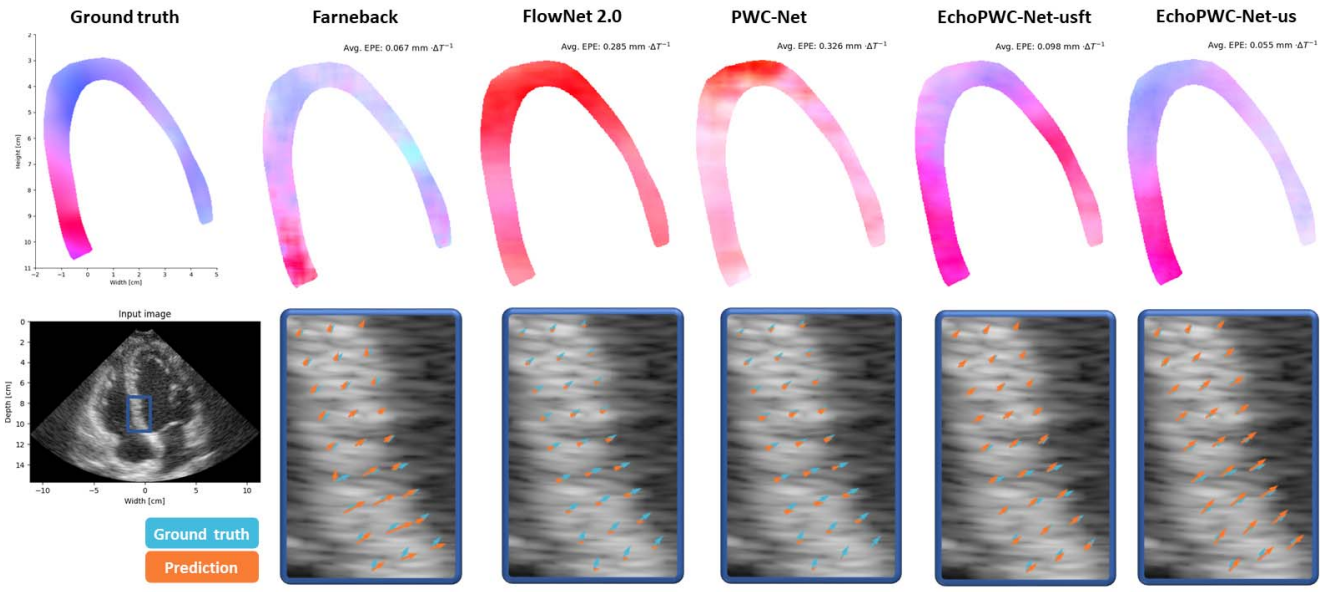
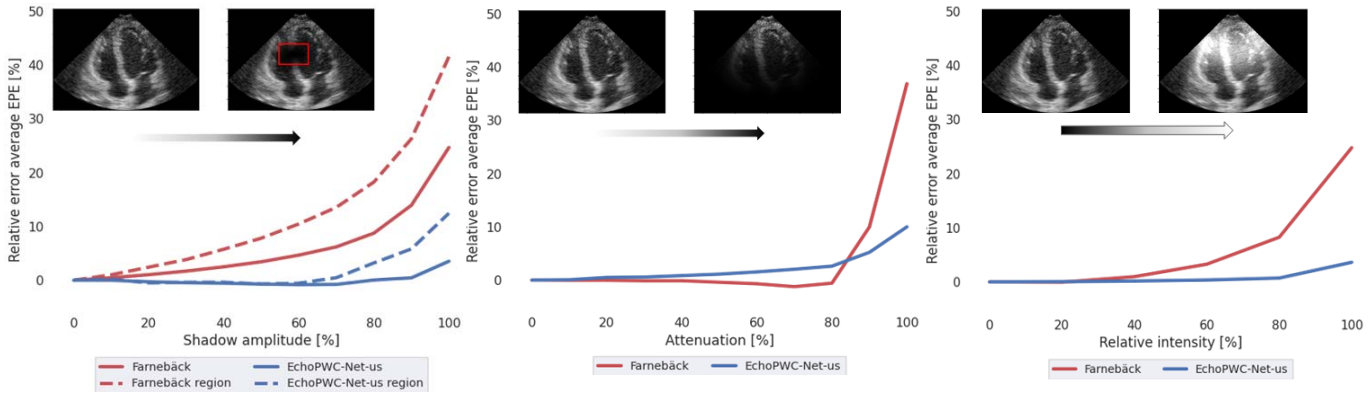


Fig. 5. Example of predicted flow patterns for the different methods within the myocardium. The upper part is color coded with hue values, color and saturation indicates direction and magnitude respectively. The average EPE is given in the upper right corner of the image of each case. The bottom part of the image shows the velocity vector comparison between ground truth and the different methods inside the base septum segment as indicated by the blue bounding box in the US image. Light blue arrows are ground truth, while orange arrows are predictions.

are reported in Table IIb and IIc respectively. A correlation plot of the regional strain estimation for one of the vendors can be seen in Fig. 4, while the plots for all vendors can be found in the supplementary material. An example of qualitative results for the different methods are shown in Fig. 5.

For reference, the correlation metrics of our implemented methods, compared to the work by Alessandrini *et al.*, is shown in Table III, and also indicated in the correlation plots. The average over all vendors is used for the different metrics.



(a) Shadow (b) Depth attenuation (c) Haze

Fig. 6. Testing of model adaptation abilities by measuring relative average end point error (EPE) as a function of fractional increase in augmentation effect. The Farnebäck method and EchoPWC-Net-us is plotted as red and blue lines respectively. (a) The shadow is applied in a specific region of the US image, as indicated by the red bounding box, and the EPE is calculated both regionally inside this box (dashed) and for the entire myocardium defined by the segmentation (solid). (b) Depth attenuation is applied with a fixed saturation area close to the base of the myocardium in radial coordinates. (c) Haze is applied to a fixed area in the upper half of the myocardium.

TABLE III
COMPARISON OF THE CONSIDERED METHODS AVERAGED OVER THE DIFFERENT VENDORS IN THE SIMULATED US DATA. METRICS INCLUDE SLOPE OF THE REGRESSION LINE α, CORRELATION COEFFICIENT ρ, BIAS μ AND 95% LIMITS OF AGREEMENT (LOA)

Method	α	ρ	μ	LOA
Alessandrini *et al.* [10]	0.55 (0.14)	0.75 (0.13)	0.37 (0.65)	6.98 (1.53)
Farnebäck	0.42 (0.15)	0.65 (0.16)	-0.16 (0.42)	7.38 (1.73)
FlowNet 2.0	0.25 (0.15)	0.49 (0.19)	-1.70 (0.86)	8.76 (1.77)
PWC-Net	0.35 (0.12)	0.58 (0.15)	0.01 (0.40)	8.01 (1.51)
PWC-Net-gray	0.24 (0.09)	0.37 (0.18)	-0.58 (0.52)	10.07 (1.94)
PWC-Net-gray-usft	0.55 (0.11)	0.66 (0.14)	-0.32 (0.54)	7.63 (1.75)
PWC-Net-us	0.57 (0.12)	0.69 (0.14)	-0.30 (0.50)	7.40 (1.25)
EchoPWC-Net	0.38 (0.25)	0.44 (0.29)	0.96 (0.74)	10.53 (4.03)
EchoPWC-Net-usft	0.48 (0.13)	0.67 (0.14)	1.11 (0.78)	7.43 (1.43)
EchoPWC-Net-us	**0.60 (0.10)**	**0.84 (0.07)**	**0.11 (0.37)**	**5.45 (1.19)**

TABLE IV
AVERAGE END POINT ERROR AND STANDARD DEVIATION (PARENTHESIS) FOR EVERY VIEW ON CLINICAL DATA COMPARED TO A COMMERCIAL METHOD

Method	A4C [mm·ΔT^{-1}]	A2C [mm·ΔT^{-1}]	APLAX [mm·ΔT^{-1}]	Average [mm·ΔT^{-1}]
Farnebäck	0.19 (0.08)	0.18 (0.08)	0.19 (0.09)	0.19 (0.08)
FlowNet 2.0	0.19 (0.08)	0.18 (0.07)	0.20 (0.09)	0.19 (0.08)
PWC-Net	0.24 (0.09)	0.24 (0.09)	0.25 (0.09)	0.24 (0.09)
PWC-Net-gray	0.25 (0.11)	0.26 (0.12)	0.27 (0.12)	0.26 (0.12)
PWC-Net-gray-usft	0.19 (0.08)	0.19 (0.08)	0.19 (0.09)	0.19 (0.08)
PWC-Net-us	0.19 (0.08)	0.18 (0.08)	0.19 (0.08)	0.19 (0.08)
EchoPWC-Net	0.23 (0.10)	0.24 (0.11)	0.26 (0.12)	0.25 (0.11)
EchoPWC-Net-usft	0.17 (0.07)	0.18 (0.07)	0.18 (0.08)	0.18 (0.07)
EchoPWC-Net-us	**0.16 (0.07)**	**0.16 (0.07)**	**0.17 (0.08)**	**0.16 (0.07)**

TABLE V
AVERAGE DIFFERENCE AND STANDARD DEVIATION (PARENTHESIS) FOR GLOBAL LONGITUDINAL STRAIN ON CLINICAL DATA COMPARING TO A COMMERCIAL METHOD

Method	A4C [%]	A2C [%]	APLAX [%]	Average [%]
Segmentation only	-0.54 (2.51)	-0.34 (3.45)	-0.03 (5.00)	-0.28 (2.36)
Tracking	-0.58 (1.79)	-0.30 (1.99)	-1.11 (2.50)	-0.71 (1.63)

Results from our model adaption study is given in Fig. 6. Here, the EchoPWC-Net-us models adaptability is measured with respect to increasing application of different augmentation effects. The relative error of average EPE as a function of specified effect is given. It is worth noting that the baseline of the two models are different, i.e. the Farnebäck method has on average a higher average EPE than EchoPWC-Net-us with no shadows. The absolute deterioration is therefore higher for Farnebäck in all cases.

B. Clinical Data

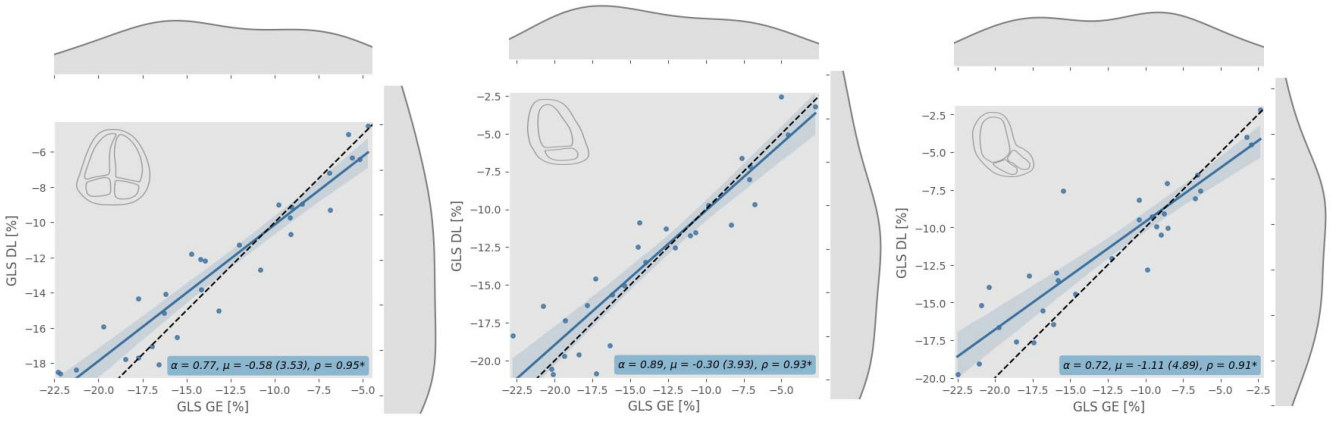
The ME methods were used on clinical in-vivo data, and compared to a commercial system by estimating the average EPE. In Table IV the results are shown for three cardiac views, and the corresponding average. We also tested the best performing model from the simulation study on this data using our pipeline for automated functional imaging. The pipeline were tested with two different flavours as specified earlier, and a summary is given in Table V. For the tracking method, a correlation plot of the GLS for each individual view is given in Fig. 7, while the average over all views is given in Fig. 8. For additional detail, Bland-Altman plots of the test data are also presented in the supplementary material.

C. Runtime Performance

The EchoPWC-Net achieves a runtime of (18.9 ± 0.7) frames per second (FPS), while the Farnebåck method can process at (8.8 ± 0.1) FPS. The frame rates of the different pipelines for estimating global longitudinal strain was (30.8 ± 0.9) FPS and (15.6 ± 0.3) FPS for segmentation and tracking respectively.

V. DISCUSSION

We have presented a method for motion estimation using DL and integrated this successfully in a pipeline for longitudinal strain measurements. The ME method is inspired by PWC-Net and relevant training strategies, but with modifications to make it more compliant for myocardial tracking. Our choices have



(a) 4CH (b) 2CH (c) APLAX

Fig. 7. Correlation plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method for specific views. Each dot represents one subject. In the bottom right corner, the slope α of the regression line, bias μ with limits of agreement (LOA) of 1.96σ in parenthesis, and correlation coefficient ρ is given.

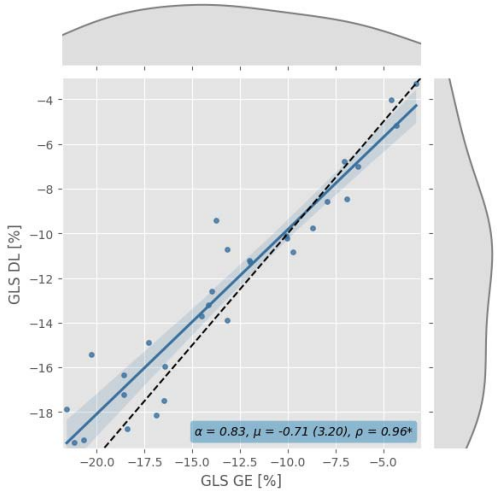


Fig. 8. Correlation plot of global longitudinal strain (GLS) estimates between commercial system and deep learning based method averaged over the three apical views. Each dot represents one subject. In the bottom right corner, the slope α of the regression line, bias μ with limits of agreement (LOA) of 1.96σ in parenthesis, and correlation coefficient ρ is given.

an intuitive motivation, firstly to increase the resemblance to echocardiography for the training data by using simulated US and relevant augmentations. Secondly, to improve the tracking task by incorporating a more direct correlation between features and loss optimization for low level feature learning, including the cost volumes at every pyramid level.

The motion magnitude of common datasets used in OF research, such as *FlyingChairs2D* and *FlyingThings3D*, is on average much higher than the displacement between frames in typical echocardiography data. This is illustrated in the supplementary material. We thus question the validity of these datasets for pretraining. As shown in Table II, training on simulated US data alone gives significantly better results. Using pretraining datasets with lower average flow could improve the results of fine tuning, but was not pursued here. We further observed a mismatch between the simulated US data and the clinical in vivo data, where the average maximum flow distribution for the latter was about twice as high. We used data augmentations to tackle this problem, but expect that an improvement of the training data quality and size will further improve the models in later iterations.

One motivation for using warping in CNNs is to mitigate the need of a large search range in the cost volume estimation. Due to the lower flow magnitudes between frames in echocardiography compared to general OF problems, incorporating the direct cost volume between features was feasible. In addition to increasing the general performance of the network for deformation imaging, the modifications introduced in EchoPWC-Net may also cause less ambiguity for occluded areas resulting from warping [35]. We believe further optimizations can be made, for instance an adaptive search range when calculating the cost volume would potentially reduce the runtime. Also the size of the pyramid can probably be reduced.

Table II suggest that the ME method producing best results is the EchoPWC-Net-us, which is trained from scratch with simulated data and several US-specific augmentation routines. Results were consistent across vendors and views. For segments, the absolute error is decreasing towards the apex, which is expected. The distribution of velocity vectors is limited for the dataset which may influence the trained models ability to generalize. Compared to the FlyingChairs dataset, the typical maximum velocity magnitude of the simulated US data is more than ten times lower. This partially explains the mismatch between the fine-tuned model and the model trained from scratch, as the latter will be biased to a lower velocity field. The qualitative results in Fig 5 further suggest the mismatch, where FlowNet 2.0 and PWC-Net yield similar results, but relatively far from the ground truth. The Farnebäck method, as well as the models trained on US data, yields good results across the entire myocardium. Noticeably, the prior has a more noisy pattern compared to the DL methods.

For strain values, the tendency of EchoPWC-Net-us is a slight underestimation for healthy segments, and a slight overestimation for sick segments. This is evident from Fig. 4, and also from the correlation plots in the supplementary material.

Noticeably, the majority of peak strain values are below 10%, and is generally low compared to clinical data. Again this indicate some limitations in the training data. A comparison to the average strain values reported by Alessandrini *et al*. [10] for the same data is given in Table III. The EchoPWC-Net-us method performs slightly better on average across vendors, especially considering the variance. Although there is potential for further improvement, these findings suggest that learning based methods can perform on par or better compared to state of the art on simulated data.

One of the major motivations of investigating the use of DL based methods, is their ability to adapt to the data representation used while training. The use of augmentation routines mimicking typical image artifacts in this scenario is therefore very appealing, as it could also address some of the big challenges with traditional methods. The result presented in Fig. 6 shows a significant improvement over a traditional OF method. For Gaussian shadowing the relative regional error is increased by less than 10% for EchoPWC-Net-us, while over 40% for the Farnebäck method. Similar effects can be seen for depth attenuation and haze application. This suggest that the ME model actively uses features from lower levels of the pyramid, and potentially the context network, in order to get the necessary global context for filling in parts of missing data. A qualitative comparison of the methods can also be found in the supplementary material. Here, the predicted flow is visualized with and without artifact for both methods. The results from the model adaption study show that the benefits of augmentation routines are twofold; in addition to increasing the effective size of the dataset, the models become more robust to image artifacts. Further studies must be conducted to evaluate the effect in vivo, but we emphasize the advantages of incorporating relevant augmentations in the training stage.

The average EPE on clinical data is significantly higher than for simulated data. We also notice that the relative improvement by training on simulated data is less effective in vivo. This may be due to the limited range of displacements present for the training data. Further, as the underlying biomechanical motion model is equal across vendors, we also suspect a slight overfit to the motion model. As can be observed in Table IV, performance improves as more distinctly relevant data is included, and the lack of relevant training data is thus believed to be a limitation which needs to be addressed.

For calculation of GLS using the pipeline we see from Table V that the measurement variance improves significantly using tracking instead of segmentation alone. The segmentation model is trained on ED and ES frames, thus calculating end-systolic strain instead of peak strain is expected to yield more similar results. We also note that the results from the APLAX view is worse than for A4C and A2C for both approaches. As the motion estimation is rather consistent across views, the myocard segmentation and the centerline extraction is the main source of this discrepancy. As shown in previous work, the segmentation performs worse on the ALAX view [24]. Also, the asymmetry of the view can complicate the centerline extraction. Better overall results can therefore be achieved by improving these components. As seen in Fig. 7 and Fig. 8 it is a significant correlation between the methods. However, we also notice a similar tendency as for simulated data, with an underestimation of larger strain values, and an overestimation of lower. The range of strain values for the DL method is therefore slightly smaller compared to the commercial method. The most probable reason for this is again the training data, where low strain values are highly over-represented [10].

In the vendor comparison study [8], the commercial system used for our in vivo data overestimates strain values by an average of 1.6% compared to the mean of all vendors. Also, software only methods from Epsilon and TomTec achieve a mean difference for GLS of $(2.50 \pm 1.94)\%$ and $(-0.70 \pm 1.68)\%$ respectively when comparing to GE. This suggests that our average difference of $(0.71 \pm 1.63)\%$ is within limits of agreement of what can be expected from different commercial systems when evaluated on the same data.

The average runtime of the networks and pipelines are reasonable compared to previously reported findings [12], [19]. The CNN is fast compared to the Farnebäck implementation used, but more work must be conducted to achieve real-time performance in echocardiography. As mentioned, we believe the network can be pruned substantially, for instance by making the search range in the cost volume adaptive for the different pyramid levels.

Although the results are encouraging, we believe there are several points that can be highlighted as a recommendation for further work in addition to what is already mentioned. Post-processing and regularization are a common part of the general workflow of strain computation [4]. This includes drift compensation, temporal and spatial smoothing, as well as state estimation or recurrent methods. In this work this was not extensively investigated, but we believe it could enhance the results if important factors are considered. Post-processing can reduce the noise, but it can also limit the range of strain values and thus reduce the ability to detect local abnormalities. Further, an investigation of regional motion patterns and strain in vivo is hard to validate, but still a direction that should be pursued to establish robust methods that allows for extended clinical use of these sensitive measurements. Representative data is the key, and continued efforts should be made to establish a larger database for training and validating motion estimation methods in echocardiography.

VI. Conclusion

In this paper we present a novel pipeline for myocardial function imaging in echocardiography using deep learning. We demonstrate that a modified PWCNet motion estimation network named EchoPWC-Net can perform on par or better compared to other known methods when training on simulated ultrasound data. Results are within limits of agreements of relevant work and commercial systems, both on in-silico and in-vivo data. We argue that the main limitations stems from limited training data, and that the results can be further improved by increased data volume, and resemblance to clinical echocardiography. Our pipeline is able to estimate longitudinal strain automatically in a prospective nature. By being simple and robust, we believe these methods can facilitate the use of deformation imaging in the clinic.

REFERENCES

[1] H. Geyer *et al.*, "Assessment of myocardial mechanics using speckle tracking echocardiography: Fundamentals and clinical applications," *J. Amer. Soc. Echocardiography*, vol. 23, no. 4, pp. 351–369, Apr. 2010.

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

[3] B. Heyde *et al.*, "Elastic image registration versus speckle tracking for 2-D myocardial motion estimation: A direct comparison *in vivo*," *IEEE Trans. Med. Imag.*, vol. 32, no. 2, pp. 449–459, Feb. 2013.

[4] M. S. Amzulescu *et al.*, "Myocardial strain imaging: Review of general principles, validation, and sources of discrepancies," *Eur. Heart J. Cardiovascular Imag.*, vol. 20, no. 6, pp. 605–619, Jun. 2019.

[5] H. Blessberger and T. Binder, "Two dimensional speckle tracking echocardiography: Basic principles," *Heart*, vol. 96, no. 9, pp. 716–722, May 2010.

[6] S. Urheim, T. Edvardsen, H. Torp, B. Angelsen, and O. A. Smiseth, "Myocardial strain by Doppler echocardiography: Validation of a new method to quantify regional myocardial function," *Circulation*, vol. 102, no. 10, pp. 1158–1164, Sep. 2000.

[7] R. M. Lang *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," *Eur. Heart J. Cardiovascular Imag.*, vol. 16, no. 3, pp. 233–271, Mar. 2015.

[8] K. Farsalinos, A. Daraban, S. Ünlü, J. Thomas, L. Badano, and J. Voigt, "Head-to-head comparison of global longitudinal strain measurements among nine different vendors: The EACVI/ASE inter-vendor comparison study," *J. Amer. Soc. Echocardiograph.*, vol. 28, no. 10, pp. 1171–1181, 2015.

[9] J.-U. Voigt *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," *Eur. Heart J. Cardiovascular Imag.*, vol. 16, no. 1, pp. 1–11, Jan. 2015.

[10] M. Alessandrini *et al.*, "Realistic vendor-specific synthetic ultrasound data for quality assurance of 2-D speckle tracking echocardiography: Simulation pipeline and open access database," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 65, no. 3, pp. 411–422, Mar. 2018.

[11] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[12] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[13] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[14] N. Duchateau, A. P. King, and M. De Craene, "Machine learning approaches for myocardial motion and deformation analysis," *Frontiers Cardiovascular Med.*, vol. 6, p. 190, Jan. 2020.

[15] 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*. Cham, Switzerland: Springer, 2018, pp. 309–316.

[16] 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*. Cham, Switzerland: Springer, 2018, pp. 21–28.

[17] 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 Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2565–2573, Dec. 2020.

[18] J. Zhang *et al.*, "Fully automated echocardiogram interpretation in clinical practice," *Circulation*, vol. 138, no. 16, pp. 1623–1635, Oct. 2018.

[19] E. Smistad *et al.*, "Real-time automatic ejection fraction and foreshortening detection using deep learning," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 67, no. 12, pp. 2595–2604, Dec. 2020.

[20] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5754–5763.

[21] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Models matter, so does training: An empirical study of CNNs for optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1408–1423, Jun. 2020.

[22] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks," *Ultrasound Med. Biol.*, vol. 45, no. 2, pp. 374–384, Feb. 2019.

[23] A. M. Fiorito, A. Ostvik, E. Smistad, S. Leclerc, O. Bernard, and L. Lovstakken, "Detection of cardiac events in echocardiography using 3D convolutional recurrent neural networks," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2018, pp. 1–4.

[24] E. Smistad, I. M. Salte, A. Ostvik, S. Leclerc, O. Bernard, and L. Lovstakken, "Segmentation of apical long axis, four-and two-chamber views using deep neural networks," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Oct. 2019, pp. 8–11.

[25] S. Leclerc *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[27] E. Smistad, A. Ostvik, B. O. Haugen, and L. Lovstakken, "2D left ventricle segmentation using deep learning," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[28] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandin. Conf. Image Anal.* Berlin, Germany: Springer, 2003, pp. 363–370.

[29] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[30] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[31] D. J. Butler *et al.*, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, A. Fitzgibbon, Eds. Berlin, Germany: Springer-Verlag, 2012, pp. 611–625.

[32] E. Smistad, K. F. Johansen, D. H. Iversen, and I. Reinertsen, "Highlighting nerves and blood vessels for ultrasound-guided axillary nerve block procedures using neural networks," *J. Med. Imag.*, vol. 5, no. 4, 2018, Art. no. 044004.

[33] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.

[34] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[35] S. Zhao, Y. Sheng, Y. Dong, E. I.-C. Chang, and Y. Xu, "MaskFlownet: Asymmetric feature matching with learnable occlusion mask," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6278–6287.