Semi-Supervised Training of Optical Flow Convolutional Neural Networks in Ultrasound Elastography*

Ali K. Z. Tehrani, Morteza Mirzaei, and Hassan Rivaz

Department of Electrical and Computer Engineering, Concordia University, Canada

Abstract. Convolutional Neural Networks (CNN) have been found to have great potential in optical flow problems thanks to an abundance of data available for training a deep network. The displacement estimation step in UltraSound Elastography (USE) can be viewed as an optical flow problem. Despite the high performance of CNNs in optical flow, they have been rarely used for USE due to unique challenges that both input and output of USE networks impose. Ultrasound data has much higher high-frequency content compared to natural images. The outputs are also drastically different, where displacement values in USE are often smooth without sharp motions or discontinuities. The general trend is currently to use pre-trained networks and fine-tune them on a small simulation ultrasound database. However, realistic ultrasound simulation is computationally expensive. Also, the simulation techniques do not model complex motions, nonlinear and frequency-dependent acoustics, and many sources of artifact in ultrasound imaging. Herein, we propose an unsupervised fine-tuning technique which enables us to employ a large unlabeled dataset for fine-tuning of a CNN optical flow network. We show that the proposed unsupervised fine-tuning method substantially improves the performance of the network and reduces the artifacts generated by networks trained on computer vision databases.

Keywords: Ultrasound Elastography · Convolutional Neural Networks (CNN) · Ultrasound-guided intervention · Unsupervised training.

1 Introduction

Ultrasound is one of the most widely used modality in medical imaging, and is the preferred modality in image-guided interventions [1–3]. UltraSound Elastography (USE) is an imaging technique which provides relative stiffness properties of the tissue, and as such, provides additional guidance during interventions. Free-hand palpation is one of the most popular methods in USE due to simplicity and availability. The basic idea of free-hand palpation is that the operator compresses the tissue by ultrasound probe, and the images before and after the compression are compared to obtain the displacement map [4]. Due to the fact that most

* Supported by NSERC Discovery Grant RGPIN 04136

compression is in the axial direction, axial displacement contains more available information than the lateral one. The axial displacement map is used to obtain the strain map, which is generally inversely proportional to the elastic modulus.

Convolutional Neural Networks (CNN) have been proven useful in optical flow estimation. Many network architectures such as FlowNet [5], FlowNet2 [6], PWC-Net [7] and LiteFlowNet [8] have been proposed. The displacement estimation step of the USE can be performed using optical flow CNNs [9–14]. However, computer vision images and ultrasound data are generally different in characteristics and the objectives. Computer vision images may contain small objects with a very different optical flow from the background (for example: a hand moves and the background is fixed). Whereas in USE, the movement is generally smooth and continuous. Another difference lies in the objective of the two tasks. The objective is to find sharp and accurate optical flows in computer vision, whereas in USE, the main goal is to obtain a differentiable displacement field. These differences led to the fact that the strain map generated by optical flow CNNs trained on computer vision images have lower bias but with higher variance compared to traditional elastography algorithms [11]. The lower bias of CNNs results in high contrast images but the high variance is amplified in the spatial differentiation step. Fine-tuning is a viable options to improve the network performance and reduce this variance [9–11, 13].

Many researchers have tried to adopt optical flow CNNs for USE using supervised fine-tuning. The general trend among researchers is to use pre-trained networks and fine-tune them with generated simulation datasets with known ground truth [10,11,13,15]. They used pre-trained well-known optical flow CNNs such as FlowNet2, PWC-Net and LiteFlowNet, and fine-tuned them using supervised techniques. The structure of the networks are also modified to address the differences of computer vision and USE in the inputs [11].

Unsupervised fine-tuning is a more appropriate option for several reasons. First, simulation techniques entail several finite element and interpolation steps, which render the accuracy of sub-pixel ground truth displacement field inaccurate. Second, the simulation database often cannot model non-linear deformation and acoustic behaviors. Last but not least, the fine-tuning may cause forgetting effects [16], if the imaging parameters of ultrasound device is not close to the simulation data to the point that the fine-tuning with simulation deteriorates results on real data [11]. By using unsupervised techniques, the network can be fine-tuned to any target domain, i.e. different ultrasound machines and different organs.

In this paper, we propose a novel unsupervised technique to fine-tune pre-trained optical flow CNNs using real ultrasound images. Our method can be considered as a form of semi-supervised learning since the pre-trained network is trained by labeled data, whereas the fine-tuning is done using data with unknown ground truth. We use LiteFlowNet [8] since it is light and has shown good performance in optical flow. However, the proposed framework can be applied to any optical flow CNN. The network estimates 2D displacements for strain values ranging from 0.5 to 5 %, the performance of the algorithm in transient

elastography [17] where displacements are very small is an area of future work. Our contribution can be summarized as follows:

1. Our results show that training on computer vision images is not enough since the statistics of ultrasound RF data and physics of the displacement field are different in these two domains.
2. We propose an unsupervised fine-tuning method in elastography.
3. We use real ultrasound images for fine-tuning, thanks to our unsupervised technique which does not need ground truth displacements.
4. We propose an automatic frame and region selection algorithm which enables the user to employ real ultrasound images without any supervision and expertise.
5. We propose a novel loss function, considering statistics of RF data and physics of the displacement field.

2 Material and Methods

2.1 Unsupervised training of optical flow networks

A critical component of unsupervised techniques is the loss function, which can be expressed as [18–21]:

$$Loss = loss_d + loss_s + loss_c \tag{1}$$

where $loss_d$ is the data loss, $loss_s$ is the smoothness loss which can also be described as smoothness regularization, and $loss_c$ is the consistency loss, which shows how different the forward and backward flows are. Data loss can be described as the difference between the first image and the warped second image. Smoothness loss controls smoothness of the displacement and usually first- and second-order derivatives are utilized for this loss [19–21]. In [18], a combination of $L1$ norm and structural similarity ($SSIM$) is employed for data loss, an edge-aware smoothness regularizer is utilized as the smoothness loss, and L_1 norm is used for consistency loss. In [19], the robust generalized Charbonnier penalty [22] of census transform [23] is employed as the data loss. The Charbonnier penalty of forward and backward displacement as consistency loss have also been exploited in [19]. Finally, this paper also used an occlusion mask to remove the occluded pixels from the loss terms to avoid back propagation of occluded regions.

2.2 Proposed Method

Inspired by other unsupervised optical flow networks, we propose a fine-tuning strategy well-adapted to USE. Let I_1 and I_2 be the first and the second images, w_f and w_b be forward ($I_1 \rightarrow I_2$) and backward flows ($I_2 \rightarrow I_1$). We define the data loss as:

$$loss_d = \left\langle \Phi(I_1 - \widetilde{I_2}) \right\rangle_{O_f} \tag{2}$$

where $\langle . \rangle_{O_f}$ is the mean of non-outlier pixels. $\widetilde{I_2}$ is the warped I_2 toward I_1 using w_f and Φ denotes Charbonnier penalty:

$$\Phi(x) = (x^2 + \varepsilon)^{\gamma} \tag{3}$$

We set γ to 0.2 similar to the fine-tuning loss in [6, 7, 11] and ε denotes a small number. There is no occlusion in USE. We borrow ideas from occlusion detection to find the outlier regions of displacement estimates and exclude them in all loss terms. In order to find outlier displacement estimates, we compare forward and backward displacement using the following equation:

$$O_f = \left| w^f + w^b \right| < \alpha \tag{4}$$

we set α empirically to 1 as we observed that in case of outlier the difference between forward and negative of the backward displacement is much larger than 1. The O_f image can be considered as a mask that selects reliable regions for the loss function. In unsupervised training, frame selection is critical since many pairs of frames are not suitable for displacement estimation due to a very large decorrelation between their ultrasound data, and grossly incorrect displacement estimates can back propagate wrong values to the network. In order to do frame selection during the training, image pairs wherein the O_f mask is 0 in more than 50% of pixels are excluded. The outlier mask can be considered as a hard threshold consistency loss.

Regarding the smoothness loss, the derivative of axial displacement is often the main concern. This derivative operation amplifies variance of the displacement estimates, reducing the contrast to noise ratio (CNR). In optimization-based elastography methods, smoothness constrains are imposed on the axial displacement [24, 25]. Here, we enforce smoothness on both displacement and its derivative. The latter is insensitive to affine deformations and performs better in the boundaries [19]. Let axial displacement be w_a^f, and a and l denote axial and lateral direction, respectively. The first order smoothing loss can be given as:

$$loss_s^1 = \lambda_1 \left\langle \Phi \left\{ \frac{\partial}{\partial a} w_a^f - < \frac{\partial}{\partial a} w_a^f > \right\} \right\rangle_{O_f} + \lambda_2 \left\langle \Phi \left\{ \frac{\partial}{\partial l} w_a^f - < \frac{\partial}{\partial l} w_a^f > \right\} \right\rangle_{O_f} \tag{5}$$

where λ_1 and λ_2 are weights associated to the axial and lateral derivative, respectively and $\langle . \rangle$ denotes mean. The average of derivatives are subtracted from the regularization term to reduce the regularization bias [11, 25]. We also consider to penalize changes in the second order axial derivative of axial displacement:

$$loss_s^2 = \lambda_3 \left\langle \Phi \left\{ \frac{\partial^2}{\partial a^2} w_a^f \right\} \right\rangle_{O_f} \tag{6}$$

The network structure for unsupervised training is shown in Fig. 1. The final loss function for training can be written as:

$$Loss = loss_d + loss_s^1 + loss_s^2 \tag{7}$$

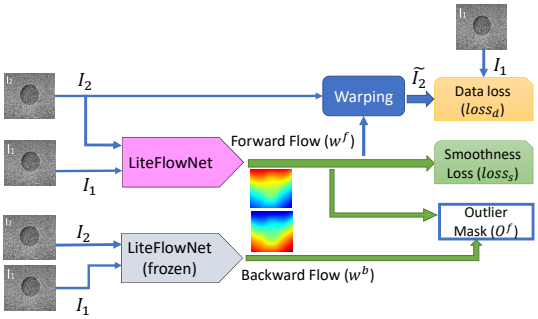


Fig. 1: Proposed network structure for unsupervised training.

It should be noted that LiteFlowNet is a multi-scale network with intermediate outputs. For training, intermediate loss and labels are required, but as suggested by [11], we only consider the last output since only small changes are required in fine-tuning.

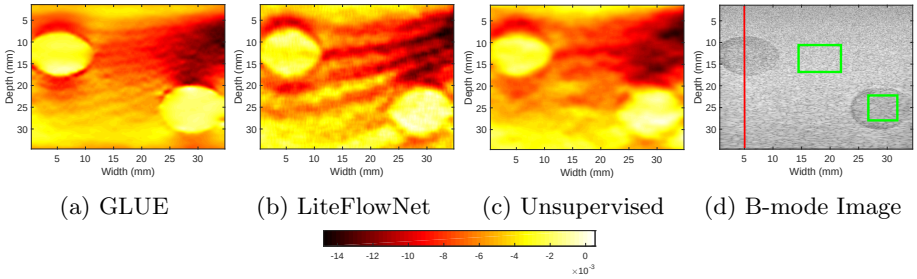
2.3 Training and Practical Considerations

Unlike computer vision images, a large input size (for example, 1920×768 in this work) is required in USE to maintain the high frequency information of the radio frequency (RF) data. This is a limiting factor for current commercial GPUs, which generally have less than 12 GB of RAM. In addition, we estimate both forward and backward displacements in our unsupervised training framework, which further intensifies the memory limitation.

To mitigate the aforementioned problem, we employ gradient checkpointing [26], where all values of forward pass are not saved into memory. When back propagation requires the values of the forward pass, they are re computed. According to [26], it decreases memory usage up to 10 times with a computational overhead of only 20-30%. Moreover, the network's weight are kept fixed in backward flow computation (the gray block in Fig. 1) to further reduce the memory usage.

Envelope of RF data along with RF data and imaginary parts of analytic signal were used as three separate channels. As suggested by [11], envelope along with RF data is used to compensate the loss of information in RF data by the downsampling steps in the network.

We used the pre-trained weights of [8], which was obtained by training on 72,000 pairs of simulated computer vision images with known optical flows. Fine-tuning was performed on 2200 pairs of real ultrasound RF data with unknown displacement maps. The test images were not seen by the network during fine-tuning. The network was trained for 20 epochs on NVIDIA TITAN V using Adam optimizer. The learning rate was set to 4e-7 and the batch size was 1 due to memory limitations. Regarding the weights of each part of the loss function, there is a trade-off between bias and variance error. Higher weights of the smoothing



(a) GLUE (b) LiteFlowNet (c) Unsupervised (d) B-mode Image

Fig. 2: Strain images of the experimental phantom with two hard inclusion. The windows used for CNR and SR computation are highlighted in the B-mode image (d). The strain of the line highlighted in red is shown in Fig. 3.

loss result in smoother but more biased results and lower weights lead to lower bias but higher variance. We empirically set λ_1, λ_2 and λ_3 to 0.5, 0.005, 0.2, as we observed that these weights had a good balance between bias and variance error.

3 Results

We validated our proposed unsupervised fine-tuning using an experimental phantom and *in vivo* data. We compare LiteFlowNet, our unsupervised fine-tuned LiteFlowNet and GLobal Ultrasound Elastography (GLUE) [24], which is a well known non-deep learning elastography method. Codes associated with all of these methods are available online.

3.1 Quantitative Metrics

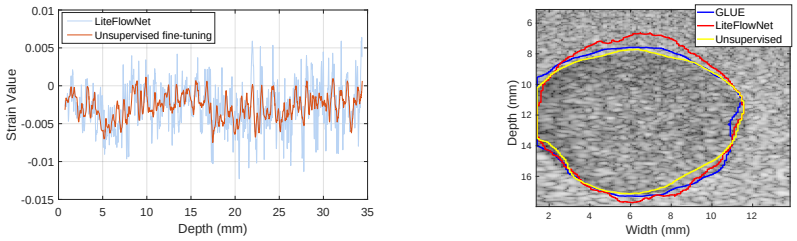
Contrast to Noise Ratio (CNR) and Strain Ratio (SR) are two popular metrics used to assess the elastography algorithms in experimental phantoms and *in vivo* data where the ground truth is unknown. These metrics are defined as [4]:

$$SR = \frac{\overline{s}_t}{\overline{s}_b}, \qquad CNR = \sqrt{\frac{2(\overline{s}_b - \overline{s}_t)^2}{{\sigma_b}^2 + {\sigma_t}^2}}, \tag{8}$$

where $\overline{s}_b$ and $\overline{s}_t$ are average values of strain in the background and target regions of the tissue, and σ_b and σ_t denote variance values of strain in the background and target regions, respectively. CNR is a proper metric to measure a combination of bias and variance error. SR sheds light on the estimator bias in real experiments wherein the ground truth strain values are unknown [11].

3.2 Experimental Phantom

We collected ultrasound images at Concordia University's PERFORM Centre using an Alpinion E-Cube R12 research ultrasound machine (Bothell, WA, USA)



(a) The strain of the line (b) The Tumor at the top left

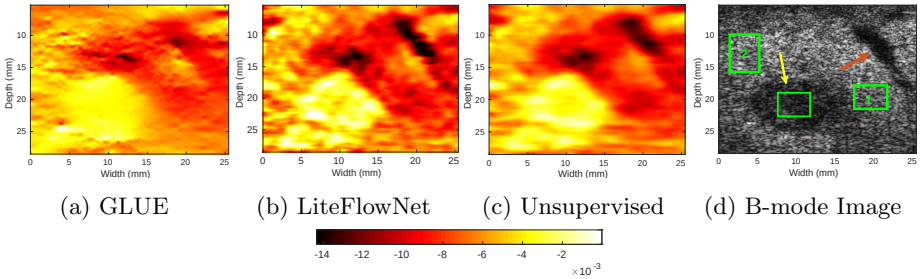
Fig. 3: The strain of the line specified in Fig. 2 using small smoothing window (a). The top left tumor with the edges obtained by the three USE methods (b).

with a L3-12H linear array at the center frequency of 10 MHz and sampling frequency of 40 MHz. A tissue mimicking breast phantom made by Zerdine (Model 059, CIRS: Tissue Simulation & Phantom Technology, Norfolk, VA) is used for data collection. The phantom contains several hard inclusions with elasticity values at least twice the elasticity of the tissue. The experimental phantom for test is from the same phantom but different part of the phantom is imaged. The composition of the phantom in test data is also different, where regions with only one inclusion are used for training and regions with two inclusions are used for testing. The test results are depicted in Fig. 2.

The unsupervised fine-tuning improves the strain quality of LiteFlowNet producing a smoother strain with less artifacts. The quantitative results are shown in Fig. 5. GLUE has the highest CNR which shows the high-quality strain but the SR is also the highest which indicates that it has the highest bias due to the strong regularization used in the algorithm. The unsupervised fine-tuning substantially improves the CNR of the network with very similar SR. In order to show the improvements better, the line specified in Fig. 2 (d) with small differentiation window is shown in Fig. 3 (a). The strain plots indicate that fine-tuning substantially reduces the variance error presented in the strain. We calculate the edges of the strain images using Canny edge detection and superimpose on the B-mode image to compare the size of different structures in in Fig. 3 (b). Here, the top left inclusion of Fig. 2 (d) is shown. It can be seen that LiteFlowNet substantially overestimates the size of the inclusion since the red curve is well outside of the inclusion. Our unsupervised fine-tuning technique corrects this overestimation.

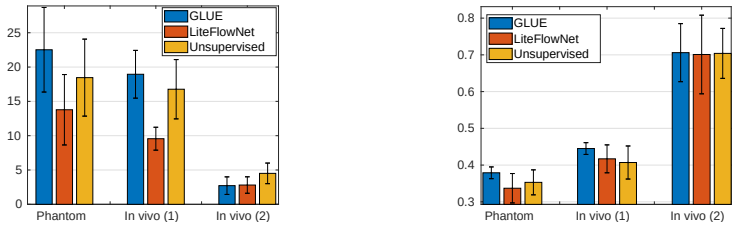
3.3 *In vivo* Data

In vivo data was collected at Johns Hopkins Hospital using a research Antares Siemens system by a VF 10-5 linear array with a sampling frequency of 40 MHz and the center frequency of 6.67 MHz. Data was obtained from patients in open-surgical RF thermal ablation for liver cancer [27]. The study was approved by the institutional review board with consent of all patients. The strains obtained



(a) GLUE (b) LiteFlowNet (c) Unsupervised (d) B-mode Image

Fig. 4: Strain images of the *in vivo* data. The two background and the target windows used for CNR and SR computation are highlighted by green in the B-mode image (d). The hard tumor and soft vein are marked by yellow and red arrows, respectively.



(a) CNR (higher is better) (b) SR (lower is better)

Fig. 5: CNR (a) and SR (b) of experimental phantom and *in vivo* data. *In vivo* (1) and *in vivo* (2) correspond to the CNR with background of 1 and 2 in Fig. 4 (d), respectively.

by the compared methods are shown in Fig. 4. GLUE (a) produce high quality strain but it is over smoothed which is evident for the vein (it is marked by the orange arrow). LiteFlowNet (b) produces a strain with many artifacts and heterogeneities inside the tumor (the tumor is marked with the yellow arrow), but it preserves the vein which is a small structure. Unsupervised fine-tuning (c) not only reduces the artifacts and heterogeneity inside the tumor, but also maintains the vein similar to LiteFlowNet.

The quantitative results are given in Fig. 5. The first background window (1 in Fig. 4 (d)) is very different than the tumor region, GLUE and the fine-tuned network produce similar CNRs between the tumor and this window. The second background (2 in Fig. 4 (d)) has very high amount of artifacts and the strain value is very close to the strain of tumor. Fine-tuned network has the best CNR for this challenging background which can be confirmed by visual assessment of Fig. 4. Regarding SR, LiteFlowNet has the lowest SR but the differences with the unsupervised fine-tuned network are negligible. Among the compared methods, GLUE has the worst SR results which indicates the high bias error presented in this method.

4 Conclusion

Herein, we proposed an semi-supervised technique for USE. We fine-tuned an optical flow network trained on computer vision images using unsupervised training. We designed a loss function suitable for our task and substantially improved the strain quality by fine-tuning the network on real ultrasound images. The proposed method can facilitate commercial adoption of USE by allowing a convenient unsupervised training technique for imaging different organs using different hardware and beamforming techniques. Inference is also very fast, facilitating the use of USE in image-guided interventions.

5 Acknowledgement

We thank NVIDIA for the donation of the GPU. The *in vivo* data was collected at Johns Hopkins Hospital. We thank E. Boctor, M. Choti and G. Hager for giving us access to this data.

References

1. S. Azizi, P. Yan, A. Tahmasebi, P. Pinto, B. Wood, J. T. Kwak, S. Xu, B. Turkbey, P. Choyke, P. Mousavi *et al.*, "Learning from noisy label statistics: detecting high grade prostate cancer in ultrasound guided biopsy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 21–29.
2. S. K. Zhou, D. Rueckert, and G. Fichtinger, *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019.
3. B. Zhuang, R. Rohling, and P. Abolmaesumi, "Region-of-interest-based closed-loop beamforming for spinal ultrasound imaging," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 66, no. 8, pp. 1266–1280, 2019.
4. J. Ophir, S. K. Alam, B. Garra, F. Kallel, E. Konofagou, T. Krouskop, and T. Varghese, "Elastography: ultrasonic estimation and imaging of the elastic properties of tissues," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 213, no. 3, pp. 203–233, 1999.
5. 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.
6. 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.
7. 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.
8. T.-W. Hui, X. Tang, and C. Change Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8981–8989.

9. 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.
10. B. Peng, Y. Xian, and J. Jiang, "A convolution neural network-based speckle tracking method for ultrasound elastography," in *2018 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2018, pp. 206–212.
11. A. K. Tehrani and H. Rivaz, "Displacement estimation in ultrasound elastography using pyramidal convolutional neural network," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2020.
12. S. Wu, Z. Gao, Z. Liu, J. Luo, H. Zhang, and S. Li, "Direct reconstruction of ultrasound elastography using an end-to-end deep neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 374–382.
13. 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*, p. 0161734620902527, 2020.
14. Z. Gao, S. Wu, Z. Liu, J. Luo, H. Zhang, M. Gong, and S. Li, "Learning the implicit strain reconstruction in ultrasound elastography using privileged information," *Medical image analysis*, vol. 58, pp. 11–18, 2019.
15. 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*, 2020.
16. Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
17. L. Sandrin, B. Fourquet, J.-M. Hasquenoph, S. Yon, C. Fournier, F. Mal, C. Christidis, M. Ziol, B. Poulet, F. Kazemi *et al.*, "Transient elastography: a new noninvasive method for assessment of hepatic fibrosis," *Ultrasound in medicine & biology*, vol. 29, no. 12, pp. 1705–1713, 2003.
18. C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.
19. S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
20. Z. Ren, J. Yan, X. Yang, A. Yuille, and H. Zha, "Unsupervised learning of optical flow with patch consistency and occlusion estimation," *Pattern Recognition*, p. 107191, 2020.
21. Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4884–4893.
22. D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.
23. R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*. Springer, 1994, pp. 151–158.
24. H. S. Hashemi and H. Rivaz, "Global time-delay estimation in ultrasound elastography," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 64, no. 10, pp. 1625–1636, 2017.

25. M. Mirzaei, A. Asif, and H. Rivaz, "Combining total variation regularization with window-based time delay estimation in ultrasound elastography," *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2744–2754, 2019.
26. T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training deep nets with sublinear memory cost," *arXiv preprint arXiv:1604.06174*, 2016.
27. H. Rivaz, E. M. Boctor, M. A. Choti, and G. D. Hager, "Real-time regularized ultrasound elastography," *IEEE transactions on medical imaging*, vol. 30, no. 4, pp. 928–945, 2011.