Cross domain knowledge distillation in realtime optical flow prediction on ultrasound sequences

No Author Given

No Institute Given

Abstract. In this paper, we propose an approach for realtime optical flow estimation in ultrasound sequences of vein and arteries based on knowledge distillation. Knowledge distillation is a technique to train a faster, smaller model by learning from cues of other models. Mobile devices with limited resources could be key in providing effective point-of-care healthcare and motivate the search of more lightweight solutions in the deep learning based image analysis. For ultrasound video analysis motion correspondences of image contents (anatomies) have to be computed for temporal context and for real time application, fast solutions are required. We use a PWC-Net's optical flow estimation output as soft targets to train a lightweight optical flow estimator. We analyse how well it works on the challenging task of fast segmentation propagation of vein and arteries in ultrasound images. Experiments show that even though we did not fine-tune the teachers on this task, a model trained with soft targets outperformed a model trained directly with labels and without a teacher.

Keywords: knowledge distillation · realtime video inference · ultrasound images

1 Introduction

The analysis of objects in a sequence of images is a task that many works have tried to tackle, most of the recent ones belonging to the deep learning field [1]. To achieve a coherent and accurate result over the different time points, it is important that the analysis of the current image considers the previous parts. One way of incorporating this temporal context is the estimation of optical flow. However, the classical methodology for the calculation is an iterative approach [2] that takes too long for realtime inference. Most recent image registration approaches based on deep learning (e.g. [3]) are computationally too expensive to be executed on mobile devices. Realtime estimation of optical flow in ultrasound sequences would be advantageous in many practical point-of-care ultrasound (POCUS) applications that are based on intelligent guidance through image analysis. The aim of this work is to train a network, that learns from larger, pre-trained flow estimation networks and is able to accurately propagate relevant information (e.g. segmentations of important anatomies) in ultrasound. Ultrasound images often exhibit ambiguous structure depiction and a network, that

employs only 2D convolution without temporal context, is not able to perfectly interpret the image with satisfying accuracy. So instead utilising the motion of the images in a network trained e.g. to propagate the anatomical labels correctly (which is usually coined weakly-supervised registration [4]) can leverage temporal context without requiring access to the whole temporal sequence. Clinically, this is relevant e.g. in the application of labeling vessels for an examination of the leg to diagnose whether a deep vein thrombosis (DVT) is present.

2 Related Work

2.1 Dynamic ultrasound analysis

The use of automated image analysis for ultrasound is constantly increasing both in research and practical clinical translations [5]. The recent MICCAI challenge CLUST [6] has studied the quality of image registration algorithms for tracking ultrasound but without realtime constraints. A Siamese network for respiratory motion estimation on ultrasound images has been proposed by Liu and colleagues [7], which is capable of tracking landmarks through a video sequence.
A system for compression-based DVT examination in ultrasound (US) images was proposed by Tanno and colleagues [8]. One of the tasks consists of classifying the compression status of a registered vein as either closed or open. The network itself uses stacked consecutive frames as input to create temporal consistency. The different task networks share the majority of convolutional layers and only separate the two tasks in the last convolutional layer, thus each task regulates the other during training.
To achieve higher temporal consistency and capture a more holistic view of dynamic sequences, optical flow estimations between frames can be leveraged. To ensure fast inference time, it is of importance that the optical flow prediction takes as little time as possible, while still generating accurate estimations.

2.2 Optical flow estimation

In recent research in deep learning and optical flow estimation numerous capable network solutions have been proposed, including Flownet [9], its evolution Flownet2 [10] and PWC-Net [**?**]. In the medical domain LapIRN [11] and PDD-Net [12] are two possible choices for estimating large deformations. Flownet uses CNN feature extractors on two images, correlates these features over a discretised displacement search window (originally 21×21 pixels with a stride of 4), and further processes these correlations to predict a flow field. Flownet2 extends the original Flownet approach by employing multiple different and fine-tuned versions of this architecture.
A more compact model, that was inspired by Flownet and Flownet2, was presented in [13]. Through pyramidal feature extraction and feature warping, together with short-range matching for larger displacements, the proposed lite Flownet outperforms the original Flownet and is as accurate as Flownet2. It

takes about 35ms on an Nvidia GTX 1080 to estimate the optical flow between two images.
PWC-Net, which was proposed by Sun et. al. [**?**], on the other hand, uses pyramidal images with a combination of a cost-volume layer and a warping layer to estimate the optical flow of the input images.
PDD-Net utilizes deformable convolution layers for feature extraction, which are then correlated. The correlation layer is followed by a min convolution and mean-field inference to predict dense displacement probabilities in volumes.
Some of these networks are larger, with up to 162m parameters and up to 0.6 seconds of inference time on an NVIDIA graphics card [9, 10]. However, these models are very accurate, which makes them valuable teachers in a student-teacher setting. Other models, such as the PDD-Net, can be compressed to use little space and computation.

2.3 Knowledge Distillation

Student-teacher learning, also known as knowledge distillation, was proposed by Li and colleagues [14]. The method uses one (or more) large and accurately trained neural network(s), also called teacher, and tries to teach the output distribution to a smaller network, also called student, by minimizing the KL divergence between the teacher's output and the students' prediction. If multiple teachers are used, the different predictions can be weighted by performance, such that they contribute accordingly to the error calculation of the student. Different variations of knowledge distillation have been explored in the last years.
An approach by Geras et. al. takes the assumption that every teacher of an ensemble of teachers approximates a different underlying function [15]. They have shown that it is possible to blend the function of an LSTM into a CNN, by using a weighted hard and soft bounding loss during training. This gives rise to the idea, that the teacher and the student architecture need not be the same, for knowledge distillation to work. This was already used in language processing, where the authors of [16] use cross-architectural knowledge distillation to improve the performance of BERT.
Yuan et. al. proposed that not only accurate teachers can be used in a knowledge distillation setting. In [17] they found that also baldly trained teachers can increase the performance of the students, as they provide a representative distribution of the classes in the classification task. Thus the teachers not only provide accurate information about the output but also provide regularized soft targets.
In [18] the authors compared the KL divergence as a loss function, which is widely used in knowledge distillation, to a mean squared error loss and found, that the mean squared error loss is superior to the KL divergence, especially, when using a small tau, as the label noise is mitigated.

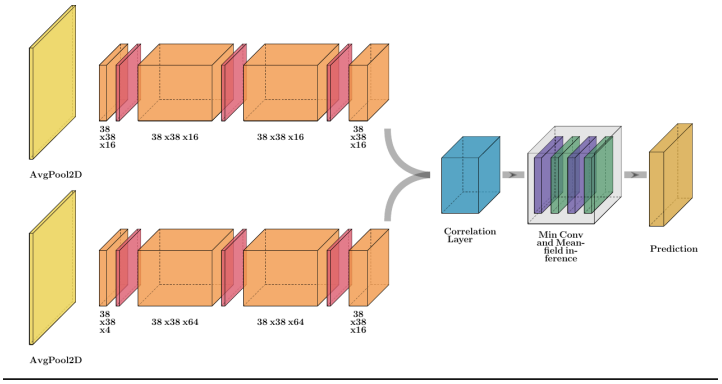

Fig. 1. Overview of the PDD-Network architecture for image registration, which comprises deformable convolutions, a correlation layer and differentiable mean-field inference as regularisation.

2.4 Contributions

We explore one of the aforementioned methods, named knowledge distillation [14] and train a small and lightweight optical flow estimator network for ultrasound motion estimation and segmentation propagation. We also compare this method to a different training setup to evaluate the usage of the distilled knowledge and find an increase in Dice score, as well as a decrease in Hausdorff distance (HD). As segmented medical reference data is scarce, this approach could potentially help increase performances for ultrasound image processing.
We aim at a short inference time of the optical flow to either create an additional input for further image analysis networks or to use the optical flow itself for segmentation propagation on mobile devices, such as tablets or phones. This constrains size and throughput of the network, as computational power on mobile devices differs greatly from stationary setups. We optimize the network for video processing and were able to generate reliable optical flow estimations in real-time.

3 Method

The PDD-Net [12] achieved a good performance in the Learn2Reg challenge [19] and a 2D implementation was made available on Kaggle [1] (1). The downsampled image is processed by three convolutional layers each followed by batch normalisation and ReLU. After a convolution, we apply an Obelisk layer [20], which is followed by two more convolutional layers. The Obelisk layer is a form of deformable convolution, which uses learnable weights and a gridsampling operator, to increase the receptive field of the next convolutional layer. The adapted

[1] https://www.kaggle.com/mattiaspaul/learn2reg-tutorial

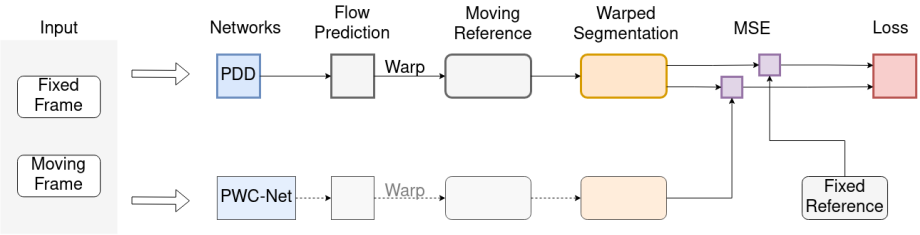

Fig. 2. Illustration of our concept for knowledge distillation for DL-based optical flow estimation. The teacher (PWC-Net) was not trained on ultrasound sequences but can provide a soft target for our student (PDD-Net) based on only a single reference frame segmentation.

weights serve as shifts in the grid, which are learned from the data. This way, the receptive field of the conv layers can be increased [20].
For a fixed and a moving image, the extracted features are correlated, akin to the Flownet-C [9] and then further processed with min convolutions and mean-field inference [21]. The whole model has an inference time of around 2.7 ms on an Nvidia RTX 2060 Ti GPU. By applying some optimization for video processing, we can reduce this time to 1.7 ms. When looking at the model (1), we can see two feature extractors, which share weights. By processing one fixed frame at time t and keeping this frame as a fixed frame, we only need to process the moving frame at point $t + x$ of the video through the CNN.
We use the established optical flow estimator PWC-Net [22] as a teacher and try to distill the knowledge into the aforementioned PDD-Net adaptation. This is done as shown in 2. We use the PWC-Net's prediction as an additional soft target during training. We calculate a weighted MSE loss between the warped moving reference segmentations of teacher and student network and include it into the regular MSE loss of the student.

Experimental setup: As the task is to propagate one reference segmentation through a whole video, we decided to create two data sets. We use a dataset of compression ultrasound sequences that are routinely performed for the diagnosis of deep vein thrombosis. We split the data set into train and evaluation IDs. From the train IDs, we first extract one data set of image pairs, that have a fixed frame distance of 6 frames. A second data set is created that is more specific to the task we want to solve. for the fine-tuning data set, we chose image pairs, where the fixed image stays the same, but the moving images vary. This leads to image pairs, that show large displacements (3), but also small displacements.

We proceed to train the PDD-Net adaptation on the training data set with additional soft targets from the PWC-Net (2) over 100 epochs with a learning rate of 0.002 and an Adam optimizer. We then trained the distilled network on the fine-tuning data set for 200 epochs with a learning rate of 0.00025. For

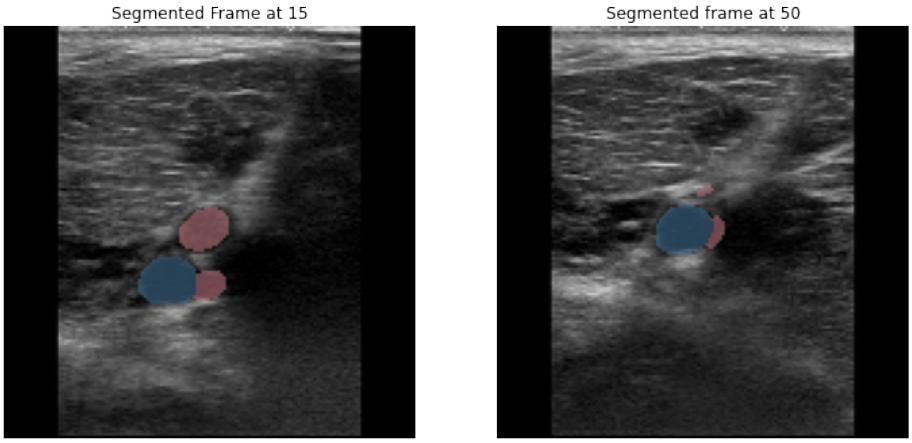

Fig. 3. Exemplary image pair used in the fine tuning data set. Reference segmentations are added for better visualization.

comparison, we also trained one version of the PDD-Net adaptation without additional soft targets.

4 Results and discussion

We evaluated the networks over 23 unseen videos containing approximately 1600 Frames overall. For each video, we selected one random frame in the first fifth of the video, which we refer to as f_t for the frame at time point t. We then took all following frames at time points $t + x$ as moving frames for the network. We estimate the optical flow between f_t and f_{t+x} and propagate the reference segmentation from f_t to f_{t+x}. This is done for all frames after time point t.
This procedure allows us to apply the mentioned runtime optimization towards video processing. By passing the fixed frame once, keeping in in memory for correlation, solely the moving frames need to be passed through the CNN for feature extraction. This reduces the inference time to 1.7 ms per image. This is about as fast, as a reference segmentation network, nnU-Net, which takes 1.6 ms on the same GPU.
As mentioned by Reinke [23] there are common limitations when applying only one metric to measure the accuracy of segmentation masks. We, therefore, evaluated the two networks with the Dice score and Hausdorff distance. The dice score is used as a measurement of overlap between the reference and predicted segmentation. It ranges from 0 to 1, where 1 is the best score, which we have denoted by $\uparrow$. The HD is used as a measurement of furthest distance between reference and predicted segmentation. We show the absolute values, where lower is better, as denoted by $\downarrow$. The mean results over all IDs can be seen in 1.
We found the distilled network to perform slightly better compared to the label

Score	registration PDD	registration PDD_{KD}	segmentation nnU-Net
vein Dice %	46.9 ± 4.13	**47.92** ± 4.15	45.93 ± 6.47
artery Dice %	44.48 ± 6.08	*46.67* ± 6.28	**66.80** ± 6.91
overall Dice %	45.69 ± 5.0	*47.3* ± 5.09	**56.36** ± 7.77
vein HD	25.28 ± 166.82	*24.16* ± 159.5	**23.71** ± 366.06
artery HD	28.3 ± 205.19	*227.7* ± 205.54	**26.88** ± 640.51
overall HD	26.79 ± 183.79	*25.93* ± 181.26	**25.33** ± 508.84

Table 1. Mean Dice ↑ over the test IDs and Mean HD ↓ over the IDs. Comparison between label loss and KD trained PDD-Nets

loss trained network over both metrics. When looking at the dice score between the two networks, we found a 2% increase in accuracy over artery segmentation and a 1% increase in vein segmentation.

When looking at the HD, we found a similar image. The KD trained network outperforms the label loss trained network slightly. We argue that this slight increase is due to the different conceptual representation learned by the distilled network, which would be in line with current research [14, 18, 16].

When compared to a segmentation network (nnU-Net [24] 1), which was trained on the same image IDs, as the optical flow estimator, we find that the distilled network is performing slightly worse in HD, and worse in Dice score. This result is somewhat expected, since the motion during longer sequences can have significant deformations (compression ultrasound of veins) and substantial drift. The frame-by-frame segmentation is in principle translation invariant and was trained with a large number of ground truth segmentation annotations. However, when visually looking at estimated segmentations (and quantitatively the variance in HD between the optical flow method and the nnU-Net), we can see that the segmentation network has limited temporal consistency. This suggests that the nnU-Net creates less smooth segmentations over a video, compared to the optical flow method. In the future, we therefore plan to experiment with the optical flow as additional input for the segmentation network.

5 Conclusion

In this paper, we presented experiments on possible benefits of cross-domain knowledge distillation (from computer vision to medical imaging) for training an optical flow estimator. By using additional teacher-generated soft targets during training, we were able to achieve a small increase in Dice score and a small decrease in Hausdorff distance. This shows that cross-domain KD can have a beneficial effect applied in the training of an image registration network.

We were able to adjust our approach to video inference, such that it is capable of running in realtime, with 1.7 ms per frame pair or more than 500 frames per second. Estimating our approach to use approximately 0.14 GFlops per image, we can calculate an upper limit of roughly 230 frames per second on modern

mobile GPUs (Qualcomm Adreno 660).
Performance of segmentation networks still exceeded segmentation via this optical flow based registration of the labels. But we suggest an increase in the segmentation networks' accuracy is possible by combining optical flow information with image features, to add temporal context to the segmentation formation. This was already suggested in previous research in medical video segmentation [25], where improved temporal coherence is reported when optical flow is incorporated. Therefore, we will further investigate the influence of optical flow on vessel segmentation in ultrasound videos.

References

1. R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, may 2020.
2. S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.
3. T. C. Mok and A. Chung, "Conditional deformable image registration with convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45, Springer, 2021.
4. Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical image analysis*, vol. 49, pp. 1–13, 2018.
5. J. A. Noble, "Reflections on ultrasound image analysis," 2016.
6. V. De Luca, J. Banerjee, A. Hallack, S. Kondo, M. Makhinya, D. Nouri, L. Royer, A. Cifor, G. Dardenne, O. Goksel, *et al.*, "Evaluation of 2d and 3d ultrasound tracking algorithms and impact on ultrasound-guided liver radiotherapy margins," *Medical physics*, vol. 45, no. 11, pp. 4986–5003, 2018.
7. F. Liu, D. Liu, J. Tian, X. Xie, X. Yang, and K. Wang, "Cascaded one-shot deformable convolutional neural networks: Developing a deep learning model for respiratory motion estimation in ultrasound sequences," *Medical Image Analysis*, vol. 65, p. 101793, 2020.
8. R. Tanno, A. Makropoulos, S. Arslan, O. Oktay, S. Mischkewitz, F. Al-Noor, J. Oppenheimer, R. Mandegaran, B. Kainz, and M. P. Heinrich, "Autodvt: Joint real-time classification for vein compressibility analysis in deep vein thrombosis ultrasound diagnostics," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 905–912, Springer, 2018.
9. 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*, pp. 2758–2766, 2015.
10. E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

11. T. C. Mok and A. C. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 211–221, Springer, 2020.
12. M. P. Heinrich, "Closing the gap between deep and conventional image registration using probabilistic dense displacement networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 50–58, Springer, 2019.
13. T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8981–8989, 2018.
14. J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.
15. K. J. Geras, A.-r. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, "Blending lstms into cnns," *arXiv preprint arXiv:1511.06433*, 2015.
16. S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, and A. Hanbury, "Improving efficient neural ranking models with cross-architecture knowledge distillation," *arXiv preprint arXiv:2010.02666*, 2020.
17. L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.
18. T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," *arXiv preprint arXiv:2105.08919*, 2021.
19. A. Hering, L. Hansen, T. C. Mok, A. Chung, H. Siebert, S. Häger, A. Lange, S. Kuckertz, S. Heldmann, W. Shao, *et al.*, "Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *arXiv preprint arXiv:2112.04489*, 2021.
20. M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical image analysis*, vol. 54, pp. 1–9, 2019.
21. P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.
22. 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*, pp. 8934–8943, 2018.
23. A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, *et al.*, "Common limitations of image processing metrics: A picture story," *arXiv preprint arXiv:2104.05642*, 2021.
24. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
25. W. Yan, Y. Wang, Z. Li, R. J. Geest, and Q. Tao, "Left ventricle segmentation via optical-flow-net from short-axis cine mri: preserving the temporal coherence of cardiac motion," in *International Conference on medical image computing and computer-assisted intervention*, pp. 613–621, Springer, 2018.